A Survey of Tagging Techniques for Music, Speech and Environmental Sound

Shufei Duan · Jinglan Zhang · Paul Roe · Michael Towsey

Abstract Sound tagging has been studied for years. Among all sound types, music, speech, and environmental sound are three hottest research areas. This survey aims to provide an overview about the state-of-the-art development in these areas. We discuss about the meaning of tagging in different sound areas at the beginning of the journey. Some examples of sound tagging applications are introduced in order to illustrate the significance of this research. Typical tagging techniques include manual, automatic, and semi-automatic approaches. After reviewing work in music, speech and environmental sound tagging, we compare them and state the research progress to date. Research gaps are identified for each research area and the common features and discriminations between three areas are discovered as well. Published datasets, tools used by researchers, and evaluation measures frequently applied in the analysis are listed. In the end, we summarise the worldwide distribution of countries dedicated to sound tagging research for years.

Keywords Sound tagging · Music tagging · Speech recognition · Environmental sound tagging · Manual tagging · Automatic tagging · Semi-automatic tagging

1 Introduction

Our life surrounds with various sounds: speech, music, animal call, aircraft, traffic, even the sound you typing words, clicking the mouse, etc. Sounds can be roughly grouped into three clusters, human voice, artificial sound, and non-artificial/natural sound. Human voice refers to sounds created by people physically such as speech, cough, and singing. Artificial sounds refer to sounds created by human activities such as traffic, aircraft, and music. Non-artificial sounds include sounds created by nature such as wind, rain, land animal, insects and marine life. These sounds make the world exclamatory and colourful. All these sounds carry information and have their own characteristics. In order to categorise different kinds of sounds and study them separately, tagging is introduced into the area of sound analysis. The act of tagging, in this context refers to the action of adding text based on metadata and annotations to specific non-textual information and data.

Initially, people classified and documented all information manually. With the development of machine technology, especially the computer science, pioneers started to research on the human-machine interaction for liberating labour force. Thanks to the great performance of automatic tagging, lots of classification work has been solved efficiently for music, speech and environmental sounds. Automatic tagging then forms the backbone of the sound recognition and classification work. However, despite the good performance, these automatic tagging machines still need information from the metadata of targets. The metadata is collected manually in several ways, social tags, survey, game, or web documents (Bertin-Mahieux et al., 2010). A basic fact also lies in that the accuracy of automatic machines still cannot catch up with the human brain. In this case, semi-automatic approaches for tagging rise which combines both manual and automatic approaches.

Among various sounds, human speech, music and environmental sounds have been studied for decades. This survey focuses on the state-of-the-art development in these three

S. Duan · J. Zhang · P. Roe · M. Towsey
Faculty of Science and Engineer, Queensland University of Technology, Brisbane, QLD, Australia
e-mail: shufei.duan@student.qut.edu.au

areas. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to the desire to automate simple tasks which necessitate human-machine interactions, speech recognition has been studied for almost 90 years since 1920s (Anusuya & Katti, 2010). Music tagging arises because people like computers help discover, manage, and describe the many new songs that become available every day (Bertin-Mahieux, et al., 2010). Since environmental sounds like bird call help ecologists monitor the environmental dynamic changes, animal calls and inner room sound tagging also draw great attention of researchers.

This paper is organised into several parts. The first part discusses the objects of this study. The second part goes through three tagging approaches for music, speech, and environmental sound. Techniques in different application areas are compared. The following part introduces published datasets for research. Research tools being used for sound analysis are introduced as well. Another part is about the worldwide contribution to sound tagging showing countries that dedicate to sound tagging analysis. The last section concludes the paper.

2 Study objects

Objects of this survey cover human speech, music, and environmental sounds.

2.1 Music

Music is an ordered arrangement of sounds and silence whose meaning is presentative rather than denotative (Clfiton, 1983). The basic features of a musical sound are pitch, intensity and timbre. In music retrieval, audio dimensions are often used for music similarity searching. The most common dimensions include: timbre, orchestration, acoustics, rhythm, melody, harmony, and structure (Orio, 2006).

2.2 Speech

Speech is the primary means of communication between humans (Furui, 2004). The properties of speech yield to language, speaker, vocabulary, speaking style (dictation or spontaneous) and speech mode (isolated or continuous) (Anusuya & Katti, 2010).

2.3 Environmental sounds

Environmental sounds include those sounds in inner room and out door. Sounds in inner rooms like meeting rooms are mainly created by human activities. While outside sounds are produced by both human and nature. Artificial sounds are due to people’s activities like aircraft and traffic. Natural sounds cover wind, rain, animal calls, insects, marine mammals, etc. The property of environmental sounds is hard to define. Probably the best feature of environmental sounds is diversity. This exists in many aspects. Take animal calls as an example. Animal calls vary according to time and season changes. Different species have different call structures. Some species have mimic behaviours. Some calls we can tell which species they belong to while some unknown call also exists (Towsey, Planitz, Nantes, Wimmer, & Roe, 2012).

3 Sound tagging

3.1 What is tagging?

(1) Music

The act of tagging, in this context refers to the action of adding text based metadata and annotations to specific non-textual information and data. (Panagakis & Kotropoulos, 2011) mentioned that “Tags are text-based labels that encode semantic information related to sound”. A tag is a keyword generated by user related with some resources (Bertin-Mahieux, et al., 2010). Automatic tagging is using machine algorithms to generate tags associated with audios. Music analysis focuses on the "identification of music genre, artist, instruments and structure" (Mitrovic, Zeppelzauer, & Breiteneder, 2006). Many songs in large music database are not tagged with semantic tags that could help users pick out the songs they want to listen to from those they do not. Auto-tagging music could help users to identify "what qualities characterize a song at a glance" and to allow users to search for the songs "most strongly by a particular word" (Hoffman, Blei, & Cook, 2009).

(2) Speech

Speech tagging focuses on the "recognition of the spoken word on syntactical level" (Mitrovic, Zeppelzauer, & Eidenberger, 2009). Automatic speech recognition is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program (Anusuya & Katti, 2010).

(3) Environmental sounds

Environmental sounds tagging is to analyse the environment that people are living, particularly animals and birds around us as people need to "study their behaviour and the way of their communication" (Franzen & Gu, 2003). Environmental sounds recognition is more complex than music analysis because environment sounds include a lot of ambient noises (R. Arora & R. Lutfi, 2009). The aim of automatic auditory scene analysis is to generate computer systems that can learn to "recognize the sound sources in a complex auditory environment" (Gunasekaran & Revathy, 2010).

3.2 Application areas

Tagging can be very advantageous when applied to particular areas such as database management and administration. Several applications and systems are involved in sound tagging technology and cover many audio fields such as animal sounds, music and human speech. Five examples of sound tagging applications are listed in Table 1.

Table 1 Samples of Sound Tagging Applications

Name	Function	Areas
Amphibulator	A device to collect wildlife environment sounds. The Amphibulator allows researchers to record audio without supervision for analysis. The system is currently being used to monitor the effects of global warming on populations of several species of amphibians in Spain and Portugal, to describe the acoustic landscape and bird populations in western Kentucky, and to study behavioural calls of midwife toads in central and northern Spain (Cambron & Bowker, 2006).	Environmental sounds
Instant Learning Sound Sensor	A context-aware system. By using this system, user is only required to input target event sounds , and it will automatically generate recognition process for small and low cost devices such as it utilises a real world sounds as rich context information without a signal processing programming (Negishi & Kawaguchi, 2007).	Environmental sounds
IPhone 4S Siri	A new virtual assistant based on voice control technology. This technology is already applied sound tagging technology into mobile device. It can recognise the human speech, complete the tasks on the speech and have conversation with human. The method is using cloud technology as the resource to recognise the speech.	Speech

SoundHound	An application of unlimited music recognition. This application is to recognise the music songs from part of the sounds. It can listen upon 10 seconds to search and then discover the sounds from a music song. This is also an excising example of application of sound tagging technologies.	Music
Shazam	A query-by-example (QBE) search service that enables users to learn the identity of audible pre-recorded music by sampling a few seconds of audio using a mobile phone as a recording device.	Music

3.3 Tagging approaches for music, speech, and environmental sounds

3.3.1 Music tagging

3.3.1.1 Manual tagging (Social tagging)

In case of music, social tags have become a significant element of music systems. There are many tasks required machines to "hear" in order to complete them, for example to discover, manage, and describe many new songs that become available every day. Social tags actually are texts generated by humans on some collaborative platforms (Bertin-Mahieux, et al., 2010). Social tags are often located within the metadata associated to an audio file. The metadata would then contain a series of textual information that represent how certain users describe a particular audio track. Furthermore, in the study of music retrieval, often the methodologies being developed with the use of social tags aim to resolve the problem known as 'cold start or tag sparsity'. This problem can simply be described as music tracks lacking the amount of tags it needs to be able to be distinguished during text based music searching. Social tags are used to categorize and retrieve contents in social tagging systems. The increasing social tagging system users not only provide information of content, but also show their preferences through tagging information. In this case, tagging information can be used in the recommender systems to make recommendations (Milicevic et al., 2010; Bischoff, Firan, Nejdl, & Paiu, 2010).

According to Last.fm (an online radio) in 2007, the types of tags can be associated with a few categories: genre, locale, mood, opinion, instrumentation, style, misc, personal, and organizational (Bertin-Mahieux, et al., 2010).

Human tags can be obtained by four sources according to (Turnbull, Barrington, Torres, & Lanckriet, 2008). These include survey, social tags, game, and web documents. Survey is the most straightforward and costly methods since people are hired to listen to sounds and tag them. However, usually there is lack of skilled people to evolve in these tasks and the cost is really high. Social tags come from human users to tag information related to the music like artist, album by using a collaborative platform. In order to reduce the cost of human tagging, different tagging games have been developed by research teams to gather clean data ((Kim, Schmidt, & Emelle, 2008); (Law, West, Mandel, Bay, & Downie, 2009); (Turnbull, et al., 2008)). Participants fill the survey because of a reward (winning), but the reward is non monetary, hence acquiring data is not as costly. The idea is to give users an incentive to apply appropriate tags to songs or song snippets. Web documents are the forth source to collect the human tags. The basic idea is to use documents available on the internet to describe audio. For instance, one could search for words that are more often associated with a particular artist than with an "average artist", and use it as a tag. One can easily gather millions of tags, but the main drawback of this method is the noise in the data (Bertin-Mahieux, et al., 2010). How to deal with noisy social tags due to people of different levels of musical knowledge remains a research problem. To reduce the noisy tags made by end users, statistical models were built to improve the accuracy. These models are specially developed for tag prediction based on the tag count information. Tags are collected through collaborative platforms such as MajorMiner game and Last.fm. By counting the number of different types of tags for the same music clip, a weight score will be added to the tag and then put into classifiers. The higher the score is, the more reliable the tag is. Through this, a tag prediction will be made and noisy tags will be reduced (Hung-Yi Lo, 2011).

3.3.1.2 Automatic tagging

Automatically extracting music information is gaining importance as a way to structure and organizes the increasingly large numbers of music files available digitally on the web (Tzanetakis & Cook, 2002). A variety of purposes can be related by using music annotations, such as searching for songs displaying special qualities, or retrieval of semantically similar songs (Coviello et al., 2010). The drawback of text based retrieval approaches is that it is impossible to search for untagged sound files (Wichern et al., 2010). To deal with this there has been recent interest in retrieving untagged audio "from text queries and the related problem of auto-tagging", for example the ability to automatically describe and tag a sound clip based on its audio content (Wichern, Yamada, Thornburg, Sugiyama, & Spanias, 2010). Consequently, there are two directions for automatic music tagging.

(1) *Tagging based on audio features.* The methodologies involving the use of audio features are modelled through the extraction of distinguishable audio tunes and patterns. Features extracted include auditory features such as loudness, pitch, brightness, bandwidth, harmonicas; musical instrument recognition features such as resonance characteristics, amplitude, envelop; human music perception such as volume levels, pitch repetition as well as the highest and lowest concord notes. Wordnet is often used as the vocabulary for matching. Multiple classifiers are applied for classification such as GMM, SVM, and AdaBoost. Representative work in this branch were published by (Wold, Blum, Keislar, & Wheaten, 1996), (Martin, 1998), (Martin, 1998), (Allegro, 2001), (McKinney & Breebaart, 2003), (Liu, 2003), (Kostek et al., 2004), (Pedro & Cano, 2005), (Eck, Lamere, Bertin-Mahieux, & Green, 2007), (Narayanan, 2007), (Burred, Cella, Peeters, Rbel, & Schwarz, 2008), (L. N. Chen, Wolfgang; Wright, Phillip., 2009), (Dhanalakshmi, Palanivel, & Ramalingam, 2009), (Edith Law, 2009), (Luke Barrington, 2009), (Lidy et al., 2010), (Lee, 2009), (Miotto, Barrington, & Lanckriet, 2010), (Kuznetsov & Pyshkin, 2010), (Gordon Wichern, 2010), (Luke Barrington, 2010), (Jun Takagi, 2011). Multiple audio features are selected and extracted for classification tasks. However, not all features can improve the accuracy. To reduce the amount of features, Eck et al. tried several methods but failed to achieve better classification results. It is clear that the result of auto-tagging does not perform better than other highly trained social tags. This leads to the question that whether it will perform better if auto-tag technique is combined with social tagging techniques (Eck, Lamere, Bertin-Mahieux, & Green, 2007).

(2) *Tagging based on the combination of social tags and audio features.* Traditional music retrieval systems often fall under the exclusive use of social tags or audio dimensions. Recent studies however, had shown that both features can be used conjunctionally and provide more significant performance towards audio classification. The system scheme usually works with weight scores or ranking systems. Both audio features and social tags are applied to different ranking systems and these ranking results are then combined to infer the tags. Related work to this area were described by (Ogihara, 2009), (Ness, Theocharis, Tzanetakis, & Martins, 2009), (Levy & Sandler, 2009), (Tingle, 2010), (Nanopoulos & Karyd is, 2011). Combination of social tags and audio features allow effective music retrieval of audio tracks with insufficient information. This help to resolve the tag sparsity problem.

Table 2 lists the most frequent features and classifiers used in music tagging. For the meaning of these features and classifiers please refer to the appendix A.

Table 2 Common features and classifiers used in music tagging

Feature	FFT, UTI, MFCC, LPC, MPEG-7, MP, SC, BW, CFRs, RS, MSC, MPCC, BIC, Roll-off, Flux, BOF, ENT, STFT, KLIEP, SCR, ZC, Entropy, LSA, SVD, Timbre, CSML, PARAFAC2, LPCC, MFCC-Delta, etc.
Classifier	HTMK, HMM, HTK, KNN, NN, Adaboost, GMAP, SML, PLR, DWCH, SVM, PLSA, GMM, CBA, VQ, MIR, BDS, KLR, IWKLR, ANN, EMD, KTN, Binary Classifier, FDA, etc.

3.3.1.3 Progress to date and research gaps

Currently, many achievements have been made towards social tagging and automatic tagging e.g. the excellent work we have listed above. Basically, there are two issues in this area: “cold start or tag sparsity” and the accuracy of automatic tagging. Though researchers have focused on these issues for years, it is still a challenge for “cold start” problem as tags are not distributed evenly. Specifically what lacks in this area is for a better weighting or filtering scheme that ignores useless social tags created by users. In other words, how to manage or evaluate the quality of social tags still needs to be studied. Likewise, for music analysis through audio features there is still the challenge of integrating more dimensions of human perception in order to better the searching solely by humming or tapping; on which case a new feature extraction methodology must be developed exclusively for this objective. Another challenge lies in the fact that large scale of data is produced by the mass online community. How to deal with this large scale of data is becoming a major problem.

3.3.2 Speech tagging

The problem is detecting, isolating and identifying the panoply of sounds that fills human every-day acoustic environment, as well as separates non-speech sounds and speech sound recognition in noisy sources (Uribe, Meana, & Miyatake, 2005). The difficulties are linguistic, cognitive boundaries, synonymy, and data scarcity, spelling errors, plurals and parts of speech. For instance, different language styles, same word with different pronunciation and tongue-tied. As speech recognition has been researched for many years, there are plenty of review papers about the state-of-the-art development in this area. Typical ones in recently two decades are (Uribe, et al., 2005) and (Anusuya & Katti, 2010). Readers can check through these papers for detail. In this section, we generally point out the main directions and branches in speech tagging.

3.3.2.1 Types of Speech Recognition

According to different types of utterances, speech recognition systems can be classified to several classes (Anusuya & Katti, 2010):

(1) *Isolated words.* Recognizers are built to accept single words at a time. Silence is required between each word or utterance happens.
(2) *Connected words.* Recognizers take separate words or utterances to be ‘run-together’ with a minimal pause between them.
(3) *Continuous speech.* Recognizers work as computer dictation, which allow users to speak almost naturally, while the system determines the content.
(4) *Spontaneous speech.* Recognizers could handle natural and not rehearsed speech. Speech features are various, such as words being run together; "ums" and "ahs", and even slight stutters.

Overall, speech recognition classification systems can be classified due to processing applications and chosen criteria.

(1) *Speech mode:* isolated speech and continuous speech.
(2) *Speaker mode:* speaker independent, speaker dependent, and speaker adaptive.
(3) *Vocabulary size:* small, medium, and large.
(4) *Speaking style:* dictation and spontaneous.

3.3.2.2 Automatic tagging (Recognition)

Techniques for automatic speech recognition have studied through three directions, acoustic phonetic approach, pattern recognition approach, and artificial intelligence approach.

(1) *Acoustic phonetic approach.* The basis of this approach is to postulate that there exist finite, distinctive phonetic units (phonemes) in spoken language and that these units are broadly characterized by a set of acoustics properties that are manifested in the speech signal over time.

(2) *Pattern recognition approach.* The pattern-matching approach involves pattern training and pattern comparison. The essential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labelled training samples via a formal training algorithm. A speech pattern representation can be in the form of a speech template or a statistical model and can be applied to a sound (smaller than a word), a word, or a phrase. In the pattern-comparison stage of the approach, a direct comparison is made between the unknown speeches (the speech to be recognized) with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of match of the patterns. Usually, pattern recognition approaches are model based, such as Hidden Markov Model (HMM), Artificial Neural Networks (ANN), Support Vector Machine (SVM), Vector Quantization (VQ) and Dynamic Time Warping (DTW).

(3) *Artificial intelligence approach.* This approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. In this, it exploits the ideas and concepts of Acoustic phonetic and pattern recognition methods. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram.

Table 3 lists the most frequent features and classifiers used in music tagging. The meaning of these features and classifiers please refer to the appendix.

Table 3 Common features and classifiers used in speech recognition

Features	STE, SC, BW, MFCC, SNR, FFT, BCI, TDMFCC, DTDMFCC, MPCC, BIC, WMFCC, LPCC, LPC, ZCR, STFT, Entropy, etc.
Classifiers	GMAP, PLP, HMM, GMM, MLP, KNN, VQ, Naive-Bayes, Decision Tree, TDNN, ASSR, DTW, IRS, SVM, MCFIS, IFIS, HTMK, HTK, ANN, LVQ, Binary Classifier, etc.

3.3.2.3 Progress to date and research gaps

Speech tagging has been studied for almost one century since 1920s. In 1994, Moore presented 20 themes which are believed to be important to the greater understanding of the nature of speech and mechanism of speech pattern processing in general. Readers can review these themes from literature (Moore, 1994). Although plenty of work has contributed to answer these questions, we are still unclear about these 20 questions so far. Speech and speaker recognition as a first step toward human-machine communication, have attracted much attention over past decades. However, we encountered a number of practical limitations which hinder a widespread deployment of application and services. What’s more, most state-of-the-art speech recognition systems make use of the acoustic signal only and ignore visual speech cues. Few studies have been applied in this area, which is supposed to be new research trend.

3.3.3 Environmental sound tagging

Environmental sound is distributed in two directions, inner room and outside. Inner room (houses and meeting-rooms) sound study aims to detect and describe human activity and to increase the robustness of automatic speech recognition systems (Temko & Nadeu, 2006). While outside sound analysis mainly focuses on wildlife monitoring and learning as they provide useful information for the environmental changing and human technology's improving. (Mitrovic, et al., 2006) mentioned that animal sounds are an area of environmental sounds that has not been investigated in detail. Recognizing sources in the environment from the sounds they create is one of the most important functions of the auditory system (Gunasekaran & Revathy, 2010). Recognizing the environment from sounds is a fundamental problem in audio processing and has significant applications in navigation and "assistive robotics and other mobile device-based services" (Chu, Narayanan, & Jay Kuo, 2008). (Weninger & Schuller, 2011) stated that in the field of bioacoustics, there is a multiplicity of approaches existing for classifying animal sound and approaches are used in order to examine "populations of certain species", for example whales or birds, thus appropriate the algorithms "to the special characteristics of animal vocalizations" involved. According to (R. Arora & R. A. Lutfi, 2009), many works have been generated in recent years to the aim of developing an automated sound recognition system that can correctly and efficiently categorize a wide variety of common environmental sounds according to their generating source.

Unlike human speech recordings and room sound recordings, which have strict constraints for recording, real world sound recordings are collected by multimedia sensors deployed in the wild environment (Cowling & Sitte, 2003), where noise is tightly constrained. Real world sound recordings are collected under unconstrained noisy conditions. Noise and variability are two issues for real world sound (Towsey, et al., 2012). Environmental acoustic recordings can obtain a wide variety of non-biological noises and a variety of animal sound. These non-biological noises have a great range of intensities and the animal sounds are affected by the physical environment (vegetation, geography etc.). Therefore, it is far more difficult for real world sound recognition than human speech and room sound recognition.

Environmental tagging refers to all non-verbal, non-communicatory sounds. Environmental tagging in this context has not been researched to the same degree of other areas of sound tagging, such as music, or speech. Fortunately, the fundamental principles and techniques used in systems designed for speech recognition and tagging can be used and applied towards environmental tagging. While speech systems aim to isolate and identify the vocalizations within the audio data, and isolate it from any background noise, the environmental recognition and annotation is the complete opposite. It is this background noise and sounds that are the main features for the system, while isolating the speech and other unwanted noise (Uribe, et al., 2005).

3.3.3.1 Manual tagging

Manual analysis provides the ability to manually inspect, play and visual acoustic recordings and associated spectrograms. It provides tools to assist in identifying vocalisations and annotating spectrograms with special tags. Manual analysis by skilled users provides an accurate and comprehensive audit of acoustic data, however processing of large volumes of data can be time consuming. The manual approach may also necessary in acoustically complex environments, where automated tools fail to discriminate between simultaneous vocalisations. Given the volume of data associated with acoustic sensing, the time and cost required to manually analyse large recording may be prohibitive (Lau et al., 2008). Additionally, these audit tasks require highly trained users experienced in identifying variations in the calls of many species. To address this issue, automatic tagging is required urgently. However, given the complexity of acoustic sensor data, fully automated analysis for a wide range of species is still a significant challenge. In this case, people start to search help from general people who can analyse data and collect data, which is known as citizen science (Truskinger et al., 2011). In many citizen science projects, participants contribute both by analysing data like *Galaxy Zoo* (http://www.galaxyzoo.org), and collecting and contributing data like *eBird* (http://www.ebird.org). One of the foremost challenges is establishing the skill level or reputation of the participant performing the collection or analysis task. To achieve this, many citizen science projects utilise reputation management to classify participants and to establish the credibility of their contributions (Truskinger, et al., 2011), (Yang, Zhang, & Roe, 2011). Even so, the accuracy and trust reliability still are big challenge in this area.

3.3.3.2 Automatic tagging

Automated acoustic analysis usually needs three steps to recognise the target: pre-processing, feature extraction and selection based on templates, and classification. Some research they add segmentation before feature extraction in order to reduce the noise affection and separate the components within one call structure (Stowell & Plumbley, 2011).

(1) *Pre-processing.* The aim of pre-processing is to expose the Acoustic Events out of Background Noise, providing clear signals for the next processing step-Feature Extraction for Classification. Signal processing techniques for noise reduction are developed according to specific applications. Basically, there are two types of applications: time domain and frequency domain. Typical work has been done by (Hu et al., 2005), (Kwan et al., 2004), (Selin, Turunen, & Tanttu, 2007). As the spectrogram is a good visualization for sound recordings, scientists turn to deal with the problem as static images. They performed noise reduction on the two dimensional (2D) sonogram but not on the audio recording (Planitz et al., 2009), (T. Brandes, Naskrecki, & Figueroa, 2006), (Agranat, 2009).

(2) *Feature extraction and selection based on templates.* According to Cowling and Sitte, feature extraction can be split into two broad types: stationary (frequency based) feature extraction and non-stationary (time-frequency based) extraction. Stationary feature extraction produces an overall result detailing the frequencies contained on the entire signal. With stationary feature extraction, no distinction is made on where these frequencies occurred in the signal. In contrast, non-stationary feature extraction splits the signal up into discrete time units. This allows frequency to be identified as occurring in a particular area of the signal, aiding understanding of the signal (Cowling & Sitte, 2003).

(3) *Classification.* The function of classification is used to identify the sound by cataloguing the features of existing sounds in some way (training) and then comparing the test sound to the database of features (testing) (Cowling & Sitte, 2003).

Acoustic event recognisers are developed according to specific application areas: inner room sound, individual target, specific species, and specific call structures.

(1) *Inner room sound.* There are groups of scientists who focus on room surroundings. They detect acoustic events in houses and meeting-rooms in order to detect and describe human activity and to increase the robustness of automatic speech recognition systems. From the year of 2006, Temko et al. have put great efforts on meeting-room acoustic event analysis (Temko & Nadeu, 2006), (Temko & Nadeu, 2009) in projects of CHIL (Computers in the Human Interaction Loop) and CLEAR (Classification of Events, Activities, and Relationships evaluation campaigns). They chose MFCCs, frequency-filtered band energies because they want to compare their discriminative capability in this application. They also chose a set of perceptual features including short-time signal energy, sub-band energies, spectral flux, zero-crossing rate and fundamental frequency after taking into account their importance and degree of interaction (Temko, Macho, & Nadeu, 2008). In terms of classifiers, they chose SVM and GMM. SVM is based on decision surfaces and GMM models data with probability distributions. After comparison between these two classifiers, SVM-based classifier outperformed GMM-based classifier.

(2)*Individual target.* In 2010, Cheng et al. chose MFCCs combined with Gaussian Mixture Model (GMM) for individual recognition of four passerines (Cheng, Sun, & Ji, 2010). According to their statement, this is the first time to combine the MFCCs with GMM for individual recognition and the results are promising. Problems are that GMM has to be improved to optimise the recognition result and large levels of background noise still are big problems for this algorithm. For wood detection, Yella et al. used acoustic analysis to test whether the existing old-structure roads, bridges and wooden railway sleepers are strong enough to be in use (Yella, Gupta, & Dougherty, 2007). This study presents a comparison of several pattern recognition techniques combined with various stationary feature extraction techniques for classification of impact acoustic emissions. Test results showed that any technique alone cannot achieve successful recognition rates.

(3) *Specific species*. Many scientists have focused on specific animal species such as frog, cane-toad, as they are very sensitive to environmental changes. In 2004, Kwan chose features of MFCCs and the classifier of GMMs to classify bird calls such as chip sparrow, Canada goose (Kwan, et al., 2004). Huang et al. used machine learning techniques for frog classification (Huang, Yang, Yang, & Chen, 2009). Hu et al. have given huge concentration on cane-road monitoring (Hu et al., 2005). They carried out the classification on the waveform of frog calls. The feature they extracted is the envelope of frog call waveform which is followed by the processing of matched filtering (Thanh, Bulusu, & Wen, 2008). However, this algorithm is not the optimal algorithm for detection and classification in general. What's more, the match templates are built in very strict conditions with no noise.

(4) *Specific call structures.* Acoustic events have different call structures. There are syllables and multi-syllables. According to the call shapes, call structures can be divided into several groups: lines, blocks, warbles, oscillations, and stacked harmonics (Duan et al., 2011). Instead of recognising specific species, scientists turn to define recognisers for special call structures as animal calls always have some similar structures**.** In 2006, Brandes et al. used techniques associated with image processing to detect and classify narrow-band cricket and frog calls (T. Brandes, et al., 2006). This is the first time to use techniques associated with image processing to spectrograms for species recognition. High true-positive accuracy can be obtained. Application can be calls with narrow-band structures. However, the accuracy largely depends on the known sonotypes and the overlap extent of the sonotype feature values. Potential of misclassification relies heavily on the extent of the libraries completeness and the known variation. In 2008, Brandes extracted peak frequency, short-time frequency and a new developed feature called contour feature vector to identify calls of cricket, frog and bird calls with frequency-modulated characters (S. T. Brandes, 2008). This method provides an effective progress on acoustic signals recognition and it achieves better results on identifying birds, crickets and frogs in a rich noise environment. Unfortunately, this method does not work well on calls with noise from wind, heavy rain and masking from large species choruses. Objects are only calls with the structure of a narrow short-time frequency bandwidth. In 2006, Chen and Maher provided an algorithm for tonal bird vocalization (harmonic or inharmonic) detection using spectral peak tracks (Z. Chen & Maher, 2006). This method has limitations in two aspects. First, the method is inappropriate for use with bird vocalizations containing periodic or noise-like components because the assumption of connected peak tracks is violated in these cases. Second, the method also is inappropriate if the underlying spectral components change too rapidly in frequency or fluctuate in amplitude such that the peak tracks cannot be determined reliably. In 2007, Selin et al. adopted wavelets in recognition of inharmonic or transient bird sounds as wavelets has ability to preserve both frequency and temporal information, and also to analyse signals which contain discontinuities and sharp spikes (Selin, et al., 2007). The limitation with this approach is that the acoustic data was chosen manually, especially for bird calls with inharmonic or transient characters. In 2009,

Bardeli et al developed an algorithm for the periodic repetition of simple elements which is often encountered in animal vocalisations (Bardeli et al., 2010). Towsey developed an oscillation detection algorithm to recognise calls that incorporate a repeating or oscillatory structure. He also developed Acoustic Event Detection (AED) to detect rectangle structures such as ground parrot call, wind and rain (Towsey, et al., 2012). Duan develop a system to detect different kinds of acoustic component such as lines, blocks in spectrograms (Duan, et al., 2011).

3.3.3.3 Semi-automatic tagging

Semi-automatic tagging provides a hybrid approach which addresses the respective strengths and weaknesses of the manual and automated techniques. Manual analysis utilises the sophisticated recognition capabilities of an expert user, but does not scale effectively for large volumes of data. Automated techniques are effective for identifying targeted species in large volume of data, however these methods require a high degree of skill to develop and are not able to cope with the variability that animal calls present. In 2011, Wimmer presented a semi-automatic tagging approach named "human-in-the-loop", which recognises that: a) many species (particularly avian species) have a broad range of vocalisations and these vocalisations may have significant regional variation; b) environmental factors such as wind, rain, vegetation and topography can attenuate, muffle and distort vocalisations considerably. Details about this model please refer to literature (Wimmer, Towsey, Planitz, Roe, & Williamson, 2010).

Table 4 lists the most frequent features and classifiers used in music tagging. The meanings of these features and classifiers are listed in appendix A.

Table 4 Common features and classifiers used in environmental sound tagging

Feature	FT, MFCC, HCC, FWT, CWT, ZCR, STE, LPC, SRF, FB, MP, DBN, mRMR, FF, HNR, MLP, LPCC, Ecology Bag, Entropy, Ceptrum feature, LoHAS, LoLAS, DSBF, FBS, SC, SS, SF, SFX, CDFs, ATFs, STE, LSTER, BP, SBC, PLP, BFCC, MFCC-Delta, MPEG-7, LLDs, FFT, STFT, etc.
Classifier	ANN, HMM, VQ, GMM, SVM, NN, SNR, DTW, Bayesian Classifiers, LDA, Decision trees, Feed forward neural network, FCDA, KNN, LSTM-RNN, RNN, LSTM, DTD, MLP, TDNN, GMM-UBM, LR-HMM, LSTM, LDA, TESPAR, AD, CQT, STS, LVQ, SOM, EDS, Binary Classifier, etc.

3.3.3.4 Progress to date and research gaps

The progress for environmental sound tagging has just started compared with the speech recognition and music tagging. The major achievement in this area is to identify a specific target. In other words, researchers build recognisers only for specific species they are interested in, such as frog, whipbird, and whale, etc. Detailed prior knowledge about the targets needs to be collected and the training data needs to be tagged and selected manually. In music tagging, we know one of the challenge is about the "cold start" problem which refers to music with no tags. Now we encounter same problem in environmental sound tagging. What we can do currently is to detect the known species, then how to detect unknown species? This is important to ecologists as it can provide information about the diversity in an area and explore new species. Another big issue is the noise. In fact, the definition of noise in environmental sound tagging is quite subjective as it depends on what signals researchers are chasing. Consequently, signal segmentation/enhancement and noise reduction also attract great attentions. However, due to the arbitrary present of noise, these systems don't work very well.

3.4 Comparison

3.4.1 Commons

In section 3.3, we introduced the state-of-the-art development and research directions of tagging techniques for music, speech, and environmental sounds. The relationship between these three research areas is quit tight. This reflects in several aspects listed below.

(1) *The act of tagging.* Though the definition for tagging in different areas has different ways to express. The core of the act of tagging is the same. Tagging is to give description to the target manually, automatically, or semi-automatically. There are two directions for tagging. One is that given the sound, users label this sound manually. This direction typically appears in the social tagging for music and environmental sounds. The other direction is that given classes of labels, assign the sound into different classes automatically or semi-automatically. This direction usually appears in speech and environmental sound recognition for identifying the source of sound. These two directions work interactively and promote each other.

(2) *Tagging techniques*. Basically, there are three tagging techniques, manually, automatically, and semi-automatically. For music and environmental sound study, these three approaches are all required as automatic methods need the metadata from people which makes human-involved approach necessary. When it comes to the speech recognition, as the ultimate goal of speech recognition is to realize the human-computer interaction efficiently and let the computer communicate with human without the language barrier, automatic recognition approach is the main technique applied.

(3) *Features and classifiers.* Although these three approaches extract different features, the basic process of automatic audio tagging is followed the same procedures shown in figure 1. The automatic tagging system could divide into four parts: the first part is audio representation, the second part is tagging data, and the third part is machine learning algorithm. Finally, the forth part is evaluation. (Bertin-Mahieux, et al., 2010) summarized that these four parts also could explain as “what audio features and tagging data it uses, what learning algorithm is used, and how performance is evaluated.” Though each area has its specific features and classifiers, we found that some features and classifiers are commonly selected and quite important for recognition work among these three application areas. Common features and classifiers are listed in the Table 5 below.

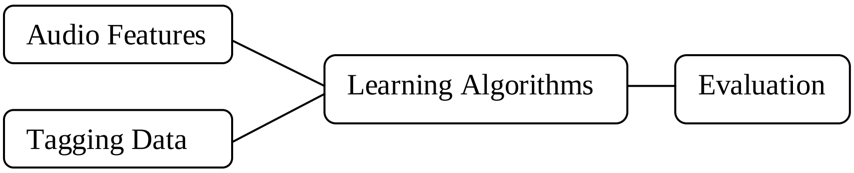

Fig. 1 The structure of a basic audio tagging system (Bertin-Mahieux, et al., 2010).

Table 5 Common features and classifiers for music, speech, and environmental sounds

Features	MFCC, FFT, LPC, LPCC, MPEG-7, SC, BW, MPCC, BIC, STE, ZCR, STFT, Entropy, MFCC-Delta
Classifiers	HMM, KNN, NN, GMAP, SVM, GMM, VQ, ANN, Binary Classifier, DTW, Bayesian Classifiers, Decision trees, MLP, TDNN, LVQ , Sonar Passive Classifier

3.4.2 Differences

Differences between music, speech and environmental sound tagging exist in many aspects.

(1) Environmental sound tagging encounters more difficulty compared with speech and music tagging due to the various noises in the data. In this case, the noise reduction or signal segmentation is a big challenge. Even the definition of noise under environmental sound tagging is hard to define. That really depends on what information users are chasing. For example, wind sound is a signal in searching for natural sound while it is noise in bird call recognition.

(2) Despite those common features and classifiers (listed in section 3.4.1), distinctive features and classifiers are also selected for specific areas. Details please refer to the tables shown in section 3.3.1 and 3.3.2. From these table, we found that the features for speech are more concentrated on MFCC, and the classifier relates to HMM and HTMK. Music has special features related to the audio dimension like MIDI, Timbre features and classifiers of Adaboost, and PLSA are quite popular. Features for environmental sound are developed for specific targets. Considering about the difficulties to identify the target under a variety of noises, quite a lot of features are extracted under certain circumstances.

3.4.3 Research Gaps

We have identified potential research gaps in section 3.3.1, 3.3.2, and 3.3.3 for music, speech, and environmental sound, respectively. To summarise, the general gaps existing in sound tagging are:

(1) Scarcity of training dataset.
(2) Lack of methods to deal with large scale of data.
(3) “Cold Start” problem for new or unknown items, especially for music and environmental sound recognition.
(4) Noise reduction, especially for environmental sound process.
(5) Lack of visual cues for acoustic signal analysis, especially for speech recognition.

4 Datasets

Many of the works reviewed above use unpublished datasets collected by the authors. Those published dataset greatly facilitate researchers’ study across the world. Thanks to the organisers of these dataset, they make the communication between different countries effective and the technique development grow quickly. Table 6 lists the datasets for sound tagging according to our review work. From this table, we can find that only few datasets are public to share with researchers. In addition, we also notice that these public datasets have been used for years. Particularly, datasets for environment sound tagging is published in 1990s by Cornell Laboratory of Ornithology. Datasets for music tagging are relatively new while speech datasets are rarely shared. Reasons for this mainly lie in the fact that the cost of data collection is very high and creators have not fully explored the datasets. However, to better develop the sound tagging research, more new and comprehensive datasets are required.

Table 6 Datasets for sound tagging

Name	Content & Feature	Application Area	Public or Not
Freesound	It is a web service which hosts a significant number of audio clips of all natures that have been uploaded by users. During the process of uploading audio to the service, the user is also required to tag	Music tagging	Yes

	the file with appropriate descriptors. Freesound also features various filters on their service which allows for the sorting of data based on its sample rate, bit depth and channels. This allows for the filtration of any “unclean” audio samples that may feature noise; though that is of little concern in music tagging. Due to Freesound’s accessibility and free use, this makes it popular as a source for datasets in the evaluation processes of music tagging systems (Takagi et al., 2011; Wichern, et al., 2010).		
CAL500	The CAL500 (Computer Audition Laboratory 500-song) is a dataset consisting of 500 audio clips, which have all been tagged by at least three individuals, using a vocabulary of 174 words and is a popular universally available dataset (Panagakis and Kotropoulos, 2011).	Music tagging	Yes
Last.fm	By the beginning of 2007, the database contained a vocabulary of 960,000 free text tags and millions of songs were annotated. Last.fm data is available through their Audioscrobbler service page (Audioscrobbler). Last.fm, provides the largest freely available collection of tagging data, but other data available from the web exist, including MusicBrainz (http://musicbrainz.org). Last.fm data have been used or described in ((Eck, et al., 2007); Lamere, 2008; (Mandel & Ellis, 2008).	Music tagging	Yes
M2VTS audio-visual database	The M2VTS audio-visual database (Dupont & Luettin, 2000) was used for all experiments. It contains 185 recordings of 37 subjects (12 females and 25 males). Each recording contains the acoustic and the video signal of the continuously pronounced French digits from zero to nine. Five recordings have been taken of each speaker, at one week intervals to account for minor face changes like beards. The video sequences consist of 286 360 pixel color images with a 25 Hz frame rate and the audio track was recorded at a 48 kHz sampling frequency and 16 bit PCM coding.	Speech recognition	Commercial Public
HU-ASA database	Weninger & Schuller (2011) mentioned that a variety of species of birds, was presented and evaluated on bird songs kept in the Animal Sound Archive of the Humboldt-University of Berlin, which will be subsequently mentioned as ‘HU-ASA database’. HU-ASA database is a	Environmental sound tagging	Yes

	large archive of animal vocalizations annotated with the species and additional metadata, including 1418 audio files available in MP3 encoding, and the total recording length of the files was 20423s (5h40min23 s). The majority of the available recordings consist of birds, mammals, 'Others' including Sauropsida, Hexapoda (Weninger & Schuller, 2011).		
Cornell-Macaulay Library of Natural Sounds	*Bird Songs of California*, Cornell Laboratory of Ornithology, Geoffrey A. Keller, 3-CD, 2003 (Stowell & Plumbley, 2011).	Environmental sound tagging	Yes
Peterson Field Guides: Bird Songs	*Western North America, A Field Guide to Western Bird Songs*, Second Ed., Cornell Laboratory of Ornithology Interactive Audio, 1992. *Eastern and Central North America, A Field Guide to Bird Songs*, Third Ed., Cornell Laboratory of Ornithology Interactive Audio, 1990.	Environmental sound tagging	Commercial Public
Common Bird Songs (Audio CD)	By Donald J. Borror, Dover Publications, 2003 *Common Birds and Their Songs* (Book and Audio CD), by Lang Elliott and Marie Read, Houghton Mifflin, 1998.	Environmental sound tagging	Commercial Public
Mitrovic's database	This database set created by (Mitrovic, et al., 2006) from an internet search. This set includes 383 samples (99 birds, 110 cats. 90 cows, 84 dogs). A sound sample contains one or more repeated sounds of an animal (such as repeated barks of a dog). Furthermore, some samples include background noise of other animals.	Environmental sound tagging	Not Sure

5 Research tools for sound tagging

Not many tools are discussed by researchers though there does exist some. Table 7 lists the common tools we reviewed in sound tagging area. Half tools or softwares are used for general sound tagging such as Weka and Matlab. Song Scope, Pamguard, Raven/XBAT softwares are developed especially for environmental sounds such as birdsongs. This is due to the various difficulties in environmental sound tagging. These tools are built to facilitate the tagging work.

Table 7 Research Tools for Sound Tagging

Name	Feature	Application Area
Wavesurfer software	It is used for visualization and manipulation. It was used for manually labelling the waveforms of the syllables in the data pre-processing stage (Selouani, Kardouchi, Hervet, & Roy, 2005).	Speech recognition
Song Scope	Song Scope is a sophisticated digital signal processing application designed to quickly and easily scan long audio recordings made in the field and automatically locate vocalizations made by specific bird species and other wildlife.	Environmental sound tagging

Pamguard	It is a marine mammal acoustic monitoring software. Pamguard provides the world standard software infrastructure for acoustic detection, localisation and classification for mitigation against harm to marine mammals, and for research into their abundance, distribution and behaviour.	Environment al sound tagging
Raven Software	Raven, produced by the Cornell Lab of Ornithology, is a software program for the acquisition, visualization, measurement, and analysis of sounds. Raven centres around audio files viewed as waveforms and spectrograms, and allows users to apply a set of analysis tools. It is designed for birdsong ananlysis workflows, so for example it provides tools to perform bandpass filters and manual or semi-automatic syllable segmentation (Stowell & Plumbley, 2011).	Environment al sound tagging especially for birdsong analysis
XBAT Software	It is also produced by the Cornell Lab of Ornithology. XBAT is similar to Raven, but it is Matlab-based, open-source (GPL), and extensible. It provides features for syllable segmentation by bandlimited power. Unlike Raven, it allows for extensibility by providing a Matlab-based API for adding filters, detectors and graphic tools (Stowell & Plumbley, 2011).	Environment al sound tagging especially for birdsong analysis
Weka	It is open source java code software created by researchers at the University of Waikato in New Zealand. It provides many different data mining and machine learning algorithms, including the following classifiers: Decision tree (j4.8, an extension of C4.5), MLP, aka multiple layer perceptron (a type of neural net), Naïve bayes, Rule induction algorithms such as JRip, Support vector machine, and many more. Weka contains modules for data preprocessing, classification, clustering and association rule extraction.	Sound tagging General tool
Matlab	Many researchers have used MATLAB to perform many of their calculations and some have used the many add in tools such as SVM-KM Toolbox which was used to conduct the IFIS and MCFIS methods (Lakshminarayanan, Raich, & Fern, 2009). Additionally the HTK toolkit can be used to claculate MFCC's as in the case of (Briggs, Raich, & Fern, 2009).	Sound tagging General tool
Sound Ruler	This tool is an open source software and is avaliable free for use. This tool is used for measuring and the graphing of sound and for teaching acoustics (Vilches, Escobar, Vallejo, & Taylor, 2006).	Sound tagging General tool
Ishmael	It is a program for acoustic analysis. It contains a spectrogram viewer, three acoustic localization methods, three methods for automatic call detection, real-time sound recording, a beamformer, and a log file annotation feature. It is more or less a collection of methods that have been found useful for analyzing acoustic data sets. Ishmael's capabilities are primarily aimed at processing large amounts of sound data quickly and relatively easily. The sound can be a collection of sound files, or a signal arriving in real time from one or more microphone(s) or hydrophone(s).	Sound tagging General tool

6 Evaluation criteria

Precision and recall are two widely used statistical criteria. Precision can be seen as a measure of exactness or fidelity, whereas recall is a measure of completeness. True Positives (TP), True Negatives (TN), False Negatives (FN) and False Positives (FP) are defined followed the definition in the paper of (Gordon et al, 2003):

(1) TP: correctly recognized positives
(2) TN: correctly recognized negatives
(3) FN: positives recognized as negatives
(4) FP: negatives recognized as positives

Precision, Recall and Accuracy are defined as (Olson et al., 2008):

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{Recall + Precision}{2}$$

The most common evaluation methods used in sound tagging area are F-score measure and Receiver operating characteristic (ROC) curves.

(1) *F-measure.* It is measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score. The F-score can be interpreted a s a weighted average of the precision and recall, where an F score reaches its best value at 1 and worst score at 0 (Yong & Ying, 2010).

$$F = \frac{2 \times Precision \times Recall}{Presicion + Recall}$$

(2) *ROC curves.* It is a graphical plot of the sensitivity (the same as recall above), or true positive rate vs. false positive rate. The ROC can also be represented equivalently by plotting the fraction of true positives out of the positives vs. the fraction of false positives out of the negatives. The ROC is also known as a Relative Operating Characteristic curve, because it is a comparison of two operating characteristics (True Positive Rate & False Positive Rate) as the criterion changes. ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution.

7 Worldwide research

In this survey, we reviewed 215 papers from the year of 1993 to 2012 for sound tagging across areas in music, speech, and environmental sounds. We chose this period mainly because of the electronic version of papers starting in 1990s. Another reason lies in the fact that we want to explore the development of sound tagging in recent 20 years, which is an important period to reflect the research trend. We have summarized countries which have the advanced technologies and have most contribution in sound tagging area. A pie chart is shown below in Figure 2 indicating the sound tagging research distributed around the world. Please note that countries presented here are according to our literature review work. Some countries may not be included due to non-comprehensive statistics.

According to this pie chart, 34 countries are involved in the sound tagging research. This number is quite promising when it comes to verify the point that sound tagging is an extremely hot research area across the world. Specifically, the United States of America holds the dominated place occupying almost 30% of whole research. Germany and Japan are in the second place for contribution to sound tagging. The third place of contributions is made by countries or areas, Taiwan, Australia, Finland, China and Spain. The percentage of countries from Germany to Canada occupies almost 40% of total. The rest countries from Switzerland to Belgium total take place of 30%.

Figure 3 shows the number of papers we reviewed for each year from 1996 to 2011. A clear signal is that sound tagging is currently a hot research area. The publications keep growing during last 16 years. Particularly, there is a significant growth in 2008 compared with year 2007. The number of publications in 2011 seems lower than that of 2010. This is because our literature review stopped in March, 2012 so some literatures of 2011 may have not been published yet. Hence, the number is not comprehensive for 2011.

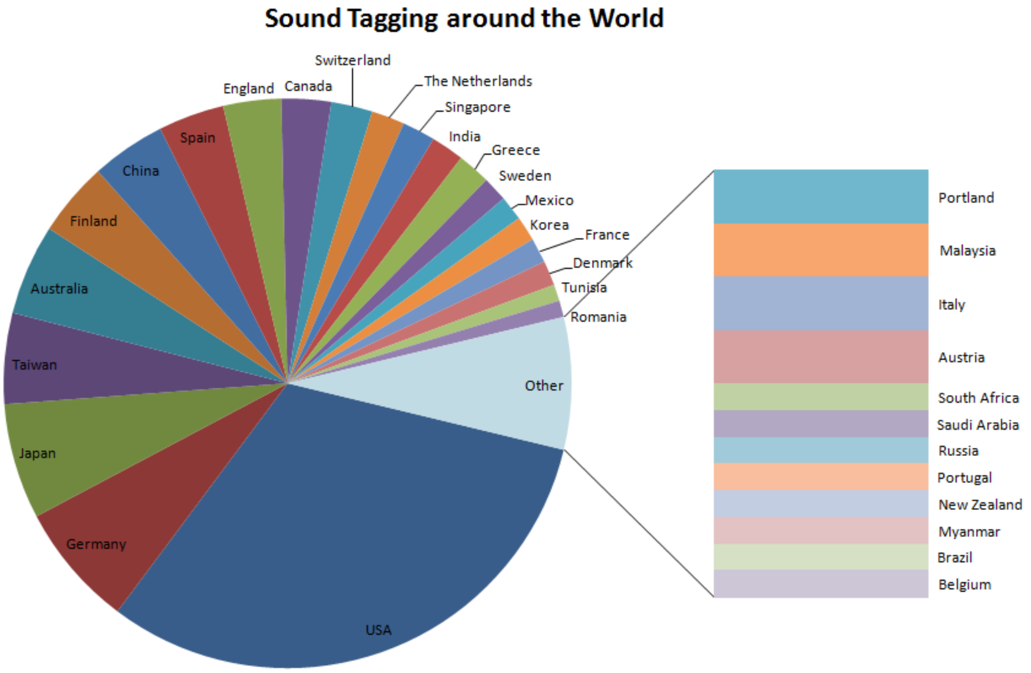

Fig. 2 Sound tagging around the world

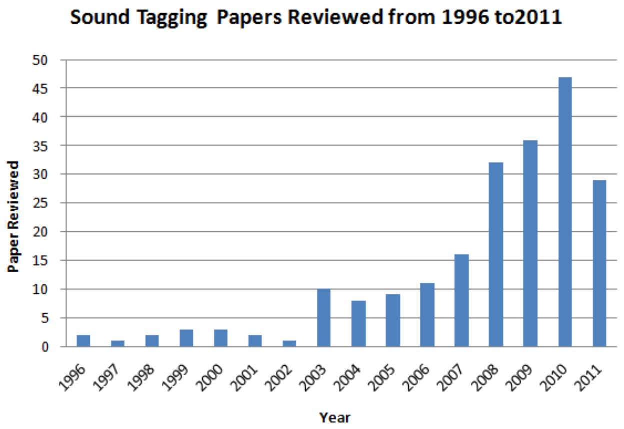

Fig. 3 Sound tagging papers reviewed from 1996 to 2011

8 Conclusion

Sound tagging has been a hot research area during last century. Considerable study and exploration have been conducted in more than 30 countries around the world. The United

States of America has played an exemplary role, holding 30% contribution to the sound tagging research. Other countries still have a long way to go towards the final automation by machine.

Three typical and interesting areas are music, speech, and environmental sound tagging. Though detailed review work has been summarised within each area, we haven’t seen a paper discussing these three subjects together. In this survey, we reviewed each direction separately and then compared them. The state-of-the-art work has been presented and potential research gaps are identified as well. For music, we found that the “cold start” problem is still a big issue and how to manage the metadata collected from social websites still needs to be addressed. Despite the great achievement for speech recognition, it is still hard and not clear to answer Moore’s 20 questions. To realise the final machine’s automation needs much more work. Speech and speaker recognition is two main branches currently. How to combine the visual feature with acoustic signal to realize the tagging work is becoming a new trend. Environmental sound tagging encounters the similar problem with music, which is lack of methods to find out the unknown species. Overall, a big issue for sound tagging is noise control. Noise reduction and signal segmentation always are the critical process for classification work.

Some of the published datasets for research were discussed. We have also surveyed the research tools used in sound tagging. To sum up, this paper has provided a survey which help new researchers who are about to start the journey with sound tagging.

References

Agranat, I. (2009). *Automatically Identifying Animal Species from their Vocalizations.* Paper presented at the Fifth International Conference on Bio-Acoustics.

Allegro, S. a. B., Michael and Launer, Stefan. (2001). Automatic sound classification inspired by auditory scene analysis. *Consistent and Reliable Acoustic Cues for Sound Analysis CRAC oneday workshop Aalborg Denmark Sunday September 2nd 2001 directly before Eurospeech 2001, 2005*, 1-4.

Anusuya, M. A., & Katti, S. K. (2010). Speech Recognition by Machine, A Review. *International Journal of Computer Science and Information Security, IJCSIS, 6*(3), 181-205.

Arora, R., & Lutfi, R. A. (2009). An efficient code for environmental sound classification. *The Journal of the Acoustical Society of America, 126*, 7.

Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K. H., & Frommolt, K. H. (2010). Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters, 31*(12), 1524-1534.

Bertin-Mahieux, T., Eck, D., Mandel, M., Mandel, M. I., Bressler, S., Shinn-Cunningham, B., et al. (2010). Automatic tagging of audio: The state-of-the-art: IGI Publishing.

Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2010). Bridging the Gap Between Tagging and Querying Vocabularies: Analyses and Applications for Enhancing Multimedia IR. *Web Semantics: Science, Services and Agents on the World Wide Web*.

Brandes, S. T. (2008). Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International, 18*(SupplementS1), S163-S173.

Brandes, T., Naskrecki, P., & Figueroa, H. (2006). Using image processing to detect and classify narrow-band cricket and frog calls. *The Journal of the Acoustical Society of America, 120*, 2950-2957.

Briggs, F., Raich, R., & Fern, X. Z. (2009, 6-9 Dec. 2009). *Audio Classification of Bird Species: A Statistical Manifold Approach.* Paper presented at the Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on.

Burred, J. J., Cella, C.-E., Peeters, G., R?bel, A., & Schwarz, D. (2008). *Using the SDIF Sound Description Interchange Format for Audio Features*. Paper presented at the ISMIR.

Cambron, M. E., & Bowker, R. G. (2006). *An Automated Digital Sound Recording System: The Amphibulator*.

Chen, L. N., Wolfgang; Wright, Phillip. (2009). Improving Music Genre Classification Using Collaborative Tagging Data.

Chen, Z., & Maher, R. C. (2006). Semi-automatic classification of bird vocalizations using spectral peak tracks. *The Journal of the Acoustical Society of America, 120*(5), 2974-2984.

Cheng, J., Sun, Y., & Ji, L. (2010). A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines. *Pattern Recognition, 43*(11), 3846-3852.

Chu, S., Narayanan, S., & Jay Kuo, C. C. (2008). *Environmental sound recognition using MP-based features*.

Coviello, E., Barrington, L., Antoni, C., & Lanckriet, G. R. G. (2010, August 9-13). *Automatic Music Tagging With Time Series Models.* Paper presented at the Proceedings of the 11th International Society for Music Information Retrieval Conference, Utrecht, Netherlands.

Cowling, M., & Sitte, R. (2003). Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters, 24*(15), 2895-2907.

Dhanalakshmi, P., Palanivel, S., & Ramalingam, V. (2009). Classification of audio signals using SVM and RBFNN. *Expert Systems with Applications, 36*(3, Part 2), 6069-6075.

Duan, S., Towsey, M., Zhang, J., Truskinger, A., Wimmer, J., & Roe, P. (2011, 6-9 Dec. 2011). *Acoustic component detection for automatic species recognition in environmental monitoring.* Paper presented at the Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2011 Seventh International Conference on.

Dupont, S., & Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *Multimedia, IEEE Transactions on, 2*(3), 141-151.

Eck, D., Lamere, P., Bertin-Mahieux, T., & Green, S. (2007). *Automatic generation of social tags for music recommendation.* Paper presented at the Advances in Neural Information Processing Systems.

Edith Law, K. W., Michael Mandel, Mert Bay J. Stephen Downie. (2009). Evaluation of Algorithms Using Games: The Case of Music Tagging. 6.

Franzen, A., & Gu, I. Y. H. (2003, 5-8 Oct. 2003). *Classification of bird species by using key song searching: a comparative study.* Paper presented at the Systems, Man and Cybernetics, 2003. IEEE International Conference on.

Furui, S. (2004). Fifty years of progress in speech and speaker recognition. *Acoustical Society of America Journal, 116*(4), 2497-2498.

Gordon, L., Chervonenkis, A. Y., Gammerman, A. J., Shahmuradov, I. A., & Solovyev, V. V. (2003). Sequence alignment kernel for recognition of promoter regions. Bioinformatics, 19(15), 1964-1971.

Gordon Wichern, M. Y., Harvey Thornburg, Masashi Sugiyama, and Andreas Spanias. (2010). Automatic Audio Tagging Using Covariate Shift Adaptation. 4.

Gunasekaran, S., & Revathy, K. (2010). *Content-based classification and retrieval of wild animal sounds using feature selection algorithm*.

Hoffman, M., Blei, D., & Cook, P. (2009). *Easy As CBA: A Simple Probabilistic Model for Tagging Music.* Paper presented at the Proc. International Symposium on Music Information Retrieval, Kobe, Japan.

Hu, W., Van Nghia, T., Bulusu, N., Chou, C. T., Jha, S., & Taylor, A. (2005, 15 April 2005). *The design and evaluation of a hybrid sensor network for cane-toad monitoring.* Paper presented at the Information Processing in Sensor Networks, 2005. IPSN 2005. Fourth International Symposium on.

Huang, C.-J., Yang, Y.-J., Yang, D.-X., & Chen, Y.-J. (2009). Frog classification using machine learning techniques. *Expert Systems with Applications, 36*(2, Part 2), 3737-3743.

Hung-Yi Lo, S.-D. L., and Hsin-Min Wang. (2011). Audio Tag Annotation and Retrieval Using Tag Count Information. 11.

Jun Takagi, Y. O., Akisato Kimura, Masashi Sugiyama, Makoto Yamada, Hirokazu Kameoka. (2011). Automatic Audio Tag Classification via Semi-supervised Canonical Density Estimation. 4.

Kim, Y. E., Schmidt, E., & Emelle, L. (2008). MoodSwings: A Collaborative Game for Music Mood Label. *ISMIR'08*, 231-236.

Kostek, B., Szczuko, P., & Zwan, P. (2004). Processing of Musical Data Employing Rough Sets and Artificial Neural Networks. *Rough Sets and Current Trends in Computing*. In S. Tsumoto, R. Slowinski, J. Komorowski & J. Grzymala-Busse (Eds.), (Vol. 3066, pp. 539-548): Springer Berlin / Heidelberg.

Kuhn, M., Wattenhofer, R., & Welten, S. (2010). *Social audio features for advanced music retrieval interfaces*. Paper presented at the Proceedings of the international conference on Multimedia.

Kuznetsov, A., & Pyshkin, E. (2010). *Searching for music: from melodies in mind to the resources on the web*. Paper presented at the Proceedings of the 13th International Conference on Humans and Computers.

Kwan, C., Mei, G., Zhao, X., Ren, Z., Xu, R., Stanford, V., et al. (2004, 17-21 May 2004). *Bird classification algorithms: theory and experimental results*. Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04).

Lakshminarayanan, B., Raich, R., & Fern, X. (2009, 13-15 Dec. 2009). *A Syllable-Level Probabilistic Framework for Bird Species Identification*. Paper presented at the Machine Learning and Applications, 2009. ICMLA '09. International Conference on.

Lau, A., Mason, R., Pham, B., Richards, M., Roe, P., & Zhang, J. (2008, 11-14 June). *Monitoring the environment through acoustics using smartphone-based sensors and 3G networking*. Paper presented at the Proceedings of the Second International Workshop on Wireless Sensor Network Deployments (WiDeploy08); 4th IEEE International Conference on Distributed Computing in Sensor Systems, DCOSS 2008, Greece.

Law, E., West, K., Mandel, M., Bay, M., & Downie, J. S. (2009). Evaluation of algorithms using games: The case of music tagging. *Evaluation*, 387-392.

Lee, J. R. a. C.-H. (2009). on the importance of modeling temporal information in music tag annotation. 4.

Levy, M., & Sandler, M. (2009). Music Information Retrieval Using Social Tags and Audio. *Multimedia, IEEE Transactions on, 11*(3), 383-395.

Lidy, T., Silla Jr, C. N., Cornelis, O., Gouyon, F., Rauber, A., Kaestner, C. A. A., et al. (2010). On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-Western and ethnic music collections. *Signal Processing, 90*(4), 1032-1048.

Liu, D. (2003). *Automatic mood detection from acoustic music data*. Paper presented at the Proceedings of the International Conference on Music Information Retrieval.

Luke Barrington, D. T., Gert Lanckriet. (2010). Auto-tagging Music Content with Semantic Multinomials.

Mandel, M. I., & Ellis, D. P. W. (2008). *Multiple-instance learning for music information retrieval* Paper presented at the the Preceedings of the 9th International Conference on Music Information Retrieval (ISMIR).

Martin, K. (1998). *Toward Automatic Sound Source Recognition: Identifying Musical Instruments*. Paper presented at the NATO Computational Hearing Advanced Study Institute.

McKinney, M. F., & Breebaart, J. (2003). *Features for audio and music classification*. Paper presented at the Proc. of the 4th ISMIR.

Milicevic, A., Nanopoulos, A., & Ivanovic, M. (2010). Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions. *Artificial Intelligence Review, 33*(3), 187-209.

Miotto, R., Barrington, L., & Lanckriet, G. (2010). *Improving Auto-tagging by Modeling Semantic Co-occurrences*. Paper presented at the International Society of Music Information Retrieval Conference, Utrecht.

Mitrovic, D., Zeppelzauer, M., & Breiteneder, C. (2006). *Discrimination and retrieval of animal sounds*.

Mitrovic, D., Zeppelzauer, M., & Eidenberger, H. (2009). *On feature selection in environmental sound recognition*. Paper presented at the ELMAR, 2009. ELMAR '09. International Symposium.

Moore, R. (1994). *Twenty things we still don't know about speech*. Paper presented at the Progress and Prospects of Speech Research and Technology: Proc. of the CRIM/FORWISS Workshop.

Nanopoulos, A., & Karydis, I. (2011, 22-27 May 2011). *Know Thy Neighbor: Combining audio features and social tags for effective music similarity*. Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.

Narayanan, S. S. a. S. (2007). Analysis of Audio Clustering Usingword Descriptions. 4.

Negishi, Y., & Kawaguchi, N. (2007). *Instant Learning Sound Sensor: Flexible Environmental Sound Recognition System*. Paper presented at the Fourth International Conference on Networked Sensing Systems.

Ness, S. R., Theocharis, A., Tzanetakis, G., & Martins, L. G. (2009). *Improving automatic music tag annotation using stacked generalization of probabilistic SVM outputs*. Paper presented at the Proceedings of the 17th ACM international conference on Multimedia.

Ogihara, F. W. X. W. B. S. T. L. a. M. (2009). Tag Integrated Multi-Label Music Style Classification with Hypergraph. *Proceedings of the 10th International Society for Music Information Retrieval Conference*, 363-368.

Olson, David, L., Delen, Dursun. (2008). Advanced Data Mining Techniques. Springer; 1 edition, page 138, ISBN 3540769161.

Orio, N. (2006). Music retrieval: a tutorial and review. *Found. Trends Inf. Retr., 1*(1), 1-96.

Panagakis, Y., & Kotropoulos, C. (2011, 22-27 May 2011). *Automatic music tagging via PARAFAC2*. Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.

Planitz, B., Roe, P., Sumitomo, J., Towsey, M., Williamson, I., Wimmer, J., et al. (2009). *Listening to nature: acoustic monitoring of the environment*. Paper presented at the Microsoft eScience Workshop.

Selin, A., Turunen, J., & Tanttu, J. T. (2007). Wavelets in recognition of bird sounds. *EURASIP J. Appl. Signal Process., 2007*(1), 141-141.

Selouani, S. A., Kardouchi, M., Hervet, E., & Roy, D. (2005, 0-0 0). *Automatic birdsong recognition based on autoregressive time-delay neural networks*. Paper presented at the Computational Intelligence Methods and Applications, 2005 ICSC Congress on.

Stowell, D., & Plumbley, M. (2011). *Birdsong and C4DM: A survey of UK birdsong and machine recognition for music researchers*: Centre for Digital Music, Queen Mary, University of London.

Studt, T. (2008). Sound and Vibration Measurement Suite, v. 6.0.

Takagi, J., Ohishi, Y., Kimura, A., Sugiyama, M., Yamada, M., & Kameoka, H. (2011, 22-27 May 2011). *Automatic audio tag classification via semi-supervised canonical density estimation*. Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.

Temko, A., & Nadeu, C. (2006). Classification of acoustic events using SVM-based clustering schemes. *Pattern Recognition, 39*(4), 682-694.

Temko, A., & Nadeu, C. (2009). Acoustic event detection in meeting-room environments. *Pattern Recognition Letters, 30*(14), 1281-1288.

Tingle, D. T., Douglass; Kim, Youngmoo. (2010). Exploring Automatic Music Annotation with “Acoustically-Objective” Tags.

Towsey, M., Planitz, B., Nantes, A., Wimmer, J., & Roe, P. (2012). A toolbox for animal call recognition. *Bioacoustics*, 1-19.

Truskinger, A. M., Yang, H., Wimmer, J., Zhang, J., Williamson, I., & Roe, P. (2011). Large scale participatory acoustic sensor data analysis : tools and reputation models to enhance effectiveness.

Turnbull, D., Barrington, L., Torres, D., & Lanckriet, G. (2008). Semantic Annotation and Retrieval of Music and Sound Effects. *Audio, Speech, and Language Processing, IEEE Transactions on, 16*(2), 467-476.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *Speech and Audio Processing, IEEE Transactions on, 10*(5), 293-302.

Uribe, O. A., Meana, H. M. P., & Miyatake, M. N. (2005, 7-9 Sept. 2005). *Environmental sounds recognition system using the speech recognition system techniques*. Paper presented at the Electrical and Electronics Engineering, 2005 2nd International Conference on.

Vilches, E., Escobar, I. A., Vallejo, E. E., & Taylor, C. E. (2006, 0-0 0). *Data Mining Applied to Acoustic Bird Species Recognition.* Paper presented at the Pattern Recognition, 2006. ICPR 2006. 18th International Conference on.

Weninger, F., & Schuller, B. Audio Recognition In The Wild: Static and Dynamic Classification on A Real-world Database of Animal Vocalizations.

Weninger, F., & Schuller, B. (2011, 22-27 May 2011). *Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations.* Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.

Wichern, G., Yamada, M., Thornburg, H., Sugiyama, M., & Spanias, A. (2010, 14-19 March 2010). *Automatic audio tagging using covariate shift adaptation.* Paper presented at the Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.

Wold, E., Blum, T., Keislar, D., & Wheaten, J. (1996). Content-based classification, search, and retrieval of audio. *Multimedia, IEEE, 3*(3), 27-36.

Yang, H., Zhang, J., & Roe, P. (2011). *Using reputation management in participatory sensing for data classification*. Paper presented at the Proeccedings of 2nd International Conference on Ambient Systems, Networks and Technologies.

Yella, S., Gupta, N. K., & Dougherty, M. S. (2007). Comparison of pattern recognition techniques for the classification of impact acoustic emissions. *Transportation Research Part C: Emerging Technologies, 15*(6), 345-360.

Yong, L., & Ying, L. (2010, 25-26 Dec. 2010). *Eco-Environmental Sound Classification Based on Matching Pursuit and Support Vector Machine.* Paper presented at the Information Engineering and Computer Science (ICIECS), 2010 2nd International Conference on.

Zhao, Z., Wang, X., Xiang, Q., Sarroff, A. M., Li, Z., & Wang, Y. (2010). *Large-scale music tag recommendation with explicit multiple attributes*. Paper presented at the Proceedings of the international conference on Multimedia.

Appendix A

Abbreviation	Descriptor	Abbreviation	Descriptor
AD	Amplitude Descriptor	LPC	Linear Prediction Coefficients
ANN	Artificial Neural Network	LSA	Latent Semantic Analysis
ATFs	Acoustic Texture Features	LSTER	Low Short-Time Energy Ratio
BCI	Brain Computer Interfaces	LSTM	Long Short-Term Memory
BFCC	Bark Frequency Cepstral Coefficients	LVQ	Learning Vector Quantization
BIC	Bayesian Information Criterion	MCFIS	Markov Chain Frame Independent Model
BOF	Bag Of Frames	MFCC	Mel-frequency Cepstral Coefficients
BP	Band Periodicity	MLP	Multi-Layer Perceptron
BW	Band Width	MP	Matching Pursuit
CBA	Codeword Bernoulli Average	MPEG-7	Moving Picture Experts Gruop
CDFs	Change Detection Features	mRMR	minimal Redundancy-Maximal Relevance
CQT	Constant Q Transform	NN	Neural Network
DTD	Data Template Detector	PARAFAC2	Parallel Factor Analysis 2
DTDMFCC	Dynamic TDMFCC	PLP	Perceptual Linear Prediction
DTW	Dynamic Time Warping	PLSA	Probabilistic Latent Semantic Analysis
EDS	Extractor Discovery System	RNN	Recurrent Neural Network
FB	Frequency Bands	RS	Rabiner and Sambur method
FFT	Fast Fourier Transform	SBC	Sub-band Based Cepstral
GMAP	Gaussian Maximum A Posteriori	SC	Spectral Centroid
GMM	Gaussian Mixture Model	SF	Spectral Flatness
HMM	Hidden Markov Model	SFX	Spectral FluX
HNR	Harmonic to Noise Ratio	SOM	Self-Organising Maps
HTK	Hidden Markov Model Toolkit	SS	Spectral Spread
IFIS	Independent Frame Independent Syllable	STE	Short Time Energy
IRS	Improved RS method	STFT	Short-Time Fourier Transform
IWKLR	Importance Weighted KLR	SVD	Singular Value Decomposition
KLR	Kernel Logistic Regression	SVM	Support Vector Machine
KNN	k-Nearest Neighbor	TDMFCC	Two-Dimensional MFCC
KTN	Know Thy Neighbor	TDNN	Time-Delay Neural Network
LDA	Linear Discriminate Analysis	UTL	Ultrasound Tagging of Light
LLDs	Low Level Descriptors	VQ	Vector Quantization
LoHAS	Length of High Amplitude Sequence	ZC	Zero Crossing
LoLAS	Length of Low Amplitude Sequence	ZCR	Zero Crossing Rate

List of Figures and Tables