



Audio Engineering Society
Convention Paper 6726

Presented at the 120th Convention
2006 May 20–23 Paris, France

This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Advanced cataloging and search techniques in audio archiving

Helge Blohmer

VCS Aktiengesellschaft, 44894 Bochum, Germany
helge.blohmer@vcs.de

ABSTRACT

Ever since the processing capabilities of computers reached the point where audio indexing and searching became possible using techniques beyond simple, manually entered textual annotation in the late 1980s, researchers have been developing such methods with varying degrees of success. Yet even today, the actual workflow in audio archives is dominated by text entry for cataloging and keywords for searching with few or none of the new methods having achieved any practical relevance. This paper is evaluating a number of techniques, both those that enhance textual retrieval and those that seek to supplant it, towards their suitability for real-world audio archiving tasks with special focus on their suitability for a short-term implementation and seamless integration into existing archive workflows.

1. INTRODUCTION

1.1. Advanced retrieval: Vision and reality

When looking back at popular scientific literature of the middle and late 1980s, one of the most defining themes was how the growing capabilities of hard- and software could be harnessed to allow computers to actually understand humans, whether this be in the fields of voice recognition, semantic analysis and text comprehension or theoretical artificial intelligence. Much of the groundbreaking work done in those days has formed the basis of many audio analysis and

processing algorithms that have come to define the industry. Some examples are:

- The concept of analyzing audio not as a waveform but as a combination of frequencies and spectra has been one of the central breakthroughs in early speech recognition approaches. Today, this concept can be found in central aspects of many compression algorithms for multimedia material in the form of the Fast Fourier Transform and the Discrete Cosine Transform.
- Psychoacoustic models have also initially been developed and embraced by researchers working on speech recognition to enable the computer to distinguish the speech from incidental noise. Today,

these models are an irreplaceable part of the MPEG audio compression algorithms.

- The basic tenets of fuzzy logic, originally an important tool for expert systems and, implicitly, neural networks, have found their way into everyday use in the form of fuzzy searching, “near match” detection and relevance rating.

The most striking realization when comparing the works of these times with today’s state of the art is however not that many of the early approaches have direct successors in today’s important multimedia-related software but that actually the tools and algorithms developed twenty years ago have made the transition into today’s industry but the underlying vision – a vision of a computer that would to a certain degree understand human language, be usable without having to comprehend formalized user interfaces and capable of giving relevant answers to human queries answering the challenge of “I need something like *this*” (with said *this* being a sample of audiovisual information) – has not become a reality. In other words, despite all advances and efforts in information organization and retrieval, the practical reality in this field is not that different from that in the 1980s. Databases are much larger, search results can be had in milliseconds instead of minutes, but the basic way of finding information inside a computer system has not changed significantly: A user starts a program, enters one or more words or other textual data items as criteria and is returned a list of items whose description – in most cases manually generated – contains these words. Yet, when looking at the scientific literature, a significant amount of progress has been made towards this vision. It has just not been translated into practical relevance.

This stark contrast between the existing powerful capabilities of the algorithms science has produced over the last two decades and their widespread unavailability has inspired the question of why this discrepancy exists and how the added capabilities can be brought into productive use for the media professional of today.

1.2. Definitions and Classification of techniques

1.2.1. The application context of audio archiving

For purposes of this paper, audio archiving is considered any process where audio material, on its own

or with accompanying other media data (images, video) is stored electronically and

- Such storage is not of the audio material only but encompasses an amount of descriptive, content-related metadata (i.e. the primary access criteria are human-relevant, not only technology-relevant),
- An important reason for the storage of this material is to enable its retrieval and reuse (i.e. it excludes collections whose purpose is preservation only)
- And the holdings are of a sufficient size and/or have sufficient regular material turnover that a user cannot be expected to have full memory of a significant part of its contents (i.e. we are not looking at small “hand archives”).

Archives meeting this definition are characterized by two major workflows, the first being the cataloging (intake of material and provision of the needed metadata) and the second being the search (locating of material based on certain content-related criteria).

Most audio archives meeting these criteria are held by broadcasters and producers of audio or multimedia material.

1.2.2. Classifying advanced archiving techniques

For purposes of evaluation in the context of the stated view of archiving, it is useful to classify and group the available technologies by two main criteria, one being the question of which of the two main workflows the technology ties into (cataloging or search) and the second being the distinction of whether the technology enhances text-based archiving workflows or supplements or replaces them with a non-text-based workflow. In addition, a third criterion affects the viability of the technologies for different use cases and that is the question whether the results of an algorithm are exact (all matches found are relevant to the query and no irrelevant matches can be generated) or heuristic (some relevant matches may be missed and/or irrelevant matches be found). Before looking at individual technologies, a short overview of the advantages and disadvantages of each variant shall be provided.

Cataloging assistance technologies are applied only once during the lifecycle of an audio item. This provides a resource advantage (processing needs to be done only

once) and possibly a time advantage (especially for algorithms that require a long running time). On the other hand, the results of these algorithms require additional storage and the output of these methods is inevitably more generic than that of a method that can take into account both the available material and the query. Also, retrofitting such techniques into existing data holdings can consume significant resources if old data is to be upgraded.

Retrieval assistance technologies provide the immediate advantage of being able to take into account both the existing data and the current query, allowing tailoring of the algorithm to each individual request. They are also useful even and especially when the existing metadata is relatively weakly structured and possibly not error-free. They can easily be applied to existing data. On the other hand, retrieval assistance is hampered by the large number of retrieval accesses and the expected short system response times, often prohibiting any algorithms requiring a large amount of processing power.

Text-aiding algorithms have the advantage of easily being integrated into existing systems. They are backwards compatible with any existing data holdings (in the case of a cataloging assistance tool, it is possible to only use it on new items and leave existing metadata untouched). They do however not add any new level of functionality to the end user side.

Non-textual algorithms are the direct result of the vision of a computer understanding human communication. They provide the user with retrieval possibilities even in cases they have no textual reference at all to start their search with. They however tend to be very processing-intensive and can in most cases only be applied to data prepared accordingly.

Exact algorithms are based on clear logical comparisons that provide "match" or "does not match" yes/no answers on each combination of query and item. They are easy to implement, however it should be noted that very few algorithms are exact beyond the purely technical meaning and rather inexact on a semantic basis. A theoretically exact word match on the word "ball" for example is a semantically inexact proposition when the query was actually intended to be for a spherical object: An object cataloged as "sphere" would be relevant but not found while "ball" in the sense of a dance would be found but not relevant.

Inexact algorithms do not provide the user with a relevance guarantee. Most however can sort their results by relevance probability (giving better matches first) and a human reviewing the list of matches is often quickly able to discern the truly relevant entries from the false hits. In spite of their disadvantages, inexact algorithms are however often useful since they can provide results on criteria or material for which there is no exact algorithm.

2. TEXT-AIDING TECHNIQUES

2.1. Text retrieval assistance

Amongst the four groups created by differentiating the possible techniques, text-aiding retrieval has the highest practical availability in commercial tools. They are also generally the easiest to implement and have found acceptance in many different situations (e.g. web page search engines, fulltext analysis) besides audio archiving. Whether in audio archiving or other applications designed at retrieval of information in large data holdings, most users will have made use of one or more of these techniques:

2.1.1. Fuzzy matching techniques

Fuzzy matching of text can generally be differentiated into two types of application:

- **Grammar-compensating fuzzy match** is a language-dependent retrieval assistance tool that exists in various levels of sophistication from the simplest varieties that only recognize words formed with the plural -s or regular verb forms ending in -s or -ed to dictionary-based solutions able to correctly match all forms of a word to their base form and vice versa even if they are completely irregular (be/is/are or eat/ate/eaten). The goal and utility of these algorithms is to enable detection of keywords in natural language, free-form descriptive text. This algorithm is most useful when archiving written works; its application to audio archiving is limited to those cases where metadata is actually written as continuous text. In most cases, such metadata will only occur when audio is linked to a textual document with its own relevance like news texts that may accompany an original sound or the composer's and editor's notes contained in an album inlay.
- **Error-compensating fuzzy match** refers to the recognition of common mistypings and misspellings.

In any professional writing scenario (and the annotation of audio material is a specialized case of professional writing), the users entering data are well-versed in spelling and grammar but time pressure and the large amount of typing required cause typographical errors. The vast majority of these errors are either letter substitutions with a letter close to it on the keyboard (“wine” might become “wibe” or “wime”) or letter swaps (“wnie” instead of “wine”). Letter omissions and additions (double keypresses) account for a lesser number of errors. Based on a dictionary approach and simple heuristics, a system can detect these misspellings and include the appropriate matches in a search result. The exactness of these algorithms is however questionable since a specific misspelling can often have multiple correct alternatives (“cit” could be “cot” or “cut”) or might even constitute legitimate words themselves (which may not be returned to avoid an overload of irrelevant results). In audio archiving, these algorithms are, despite their wide availability, thus of a relatively low relevance.

As a general evaluation, it should thus be said that the relevance and benefit of fuzzy-matching techniques of words on audio archiving is not significant. Even though they are probably the most wide-spread inclusions in commercial systems thanks to their ease of implementation, their utility is confined to a narrow application area.

2.1.2. Synonym-based searching / thesaurus

This technique is related to the previous ones in that its implementation is less of a complex algorithm but rather based on querying a dictionary. It seeks to eliminate the problem of a researcher using a different word for the same concept or object than the person who originally did the annotation. This problem occurs in nearly all archives – the only way to prevent it is the use of a strictly controlled vocabulary both on annotation and retrieval. Since very few archives still work under the premise that retrieval is always done by a specifically trained professional (archivist), this means that the problem is essentially universal. Yet, synonym searches have found a very low acceptance in practice for two reasons:

- The implementation of a full-fledged thesaurus is an extremely time-consuming task since it essentially constitutes the writing of a complete dictionary covering the meaning of thousands of words. In

addition, this task needs to be redone for each and every language the software is released in, a special problem considering the international operation of most media technology providers.

- Synonym searches tend to provide a large number of false negatives. Since the synonyms for words have subtle meaning differences from the original word – and sometimes, if the word has multiple meanings, can even mean something completely unintended – the results of an indiscriminating synonym search are inherently overloaded with false positives. This problem is compounded if multiple words are searched for at the same time.

Thesaurus techniques have the potential to be immediately useful to archive workflows and enhance productivity if implemented correctly. The major obstacles towards a successful use of this technology can be overcome relatively easily:

- The main development and cost issue in providing a synonym search is the creation of the thesaurus itself. This work can be done by specialized companies who provide the data basis in a library format. Measures need to be undertaken to ensure the integrity of the data since copyright violations on such lists would be hard to prove. For major languages, thesaurus components are available ready for integration today, however few vendors offer more than just a few languages at a time, making thesauri difficult to integrate in a finished product as is.
- On the usage side, thesauri need one additional step in the search process to be successful. Whenever a word is searched for with synonyms enabled, the user needs to be presented with a list of possible synonyms from which he then selects the relevant choices. This becomes impractical when more than two terms are being clarified this way, thus the user must decide which terms to expand on and which not. The best workflow here is to first present the exact results and, in a header or footer offer the possibility to expand terms to include selected synonyms.

In summary, it is not practical for single vendors to make thesaurus searching a default feature of their product, yet the technique offers a lot of reward and in conjunction with specialized vendors and/or industry cooperations, offers one of the best return on investment ratios of all techniques discussed.

2.1.3. Relevance evaluation schemas

While grouped in this section, relevance evaluation can be applied to any retrieval tool, text-based or not. It uses various algorithms based on search token frequency, token placement and comparison to average frequencies to derive a measurement that indicates probability of a specific result being relevant. Some sample approaches include:

- When indexing free-form text, relevance is considered higher for texts containing multiple copies of a word searched for than those containing it only once.
- When searching on fixed keywords, items where the searched-for keyword appears in first position are preferred over those showing it in a later position.
- When using inexact matching technologies (e.g. searching on the result of speech-to-text algorithms), better matches are presented first.
- A query by humming system assigns a match by pitch a higher priority than a match by rhythm.

Relevance evaluation is never used by itself but always in conjunction with other techniques. It is very easily integrated into the workflow since it does not change anything in how the product is used except that the user is likely to more quickly find desired materials in the result. However, relevance prediction is in itself a rather inexact science and the underlying algorithms represent a substantial development effort with a significant benefit being realized only when result lists are longer than approximately 10 to 20 items. The use of good relevance evaluation fitting the searching techniques used can be a major selling point to a manufacturer of large archiving systems.

2.2. Textual annotation assistance

This group of techniques encompasses those algorithms that improve the quality or workflow of generating the textual metadata accompanying an audio object. They range from very simple analysis to provide accurate information about the technical details of an audio item to complex audio analysis tools creating content-based information.

2.2.1. Formal media evaluation

Many audio archiving systems already provide tools of a varying nature to relieve the ingest operator from entering the technical metadata, i.e. all data not describing the semantic content but items that can be derived directly from the sound file with relatively simple measurements. A far from exhaustive list of these criteria includes:

- Audio duration
- Number and configuration of channels
- File format
- Peak and average levels
- Signal to noise ratio
- Location of audio breaks (silent intervals)

While all of these criteria are useful in evaluating the relevance of a search hit, they are rarely if ever used in searches – if used at all, they make restricting criteria like “only return matches that are correctly normalized to -9 dB and have a base noise level of -75 dB or better”. More advanced algorithms in this section have a limited applicability in music cataloging, allowing the computer to detect the key and beats per minute of a musical work. These concepts have found a niche in archives for DJs and music rotations, allowing a music editor to find the required raw material for samplers. Techno mixes and popular music programming which is also most likely to remain the extent of their usability.

2.2.2. Collective cataloging efforts

While not particularly advanced in technology, the idea of sharing cataloging work among not just the staff of one institution but that of many, often including private citizens as well, is of a strong economic relevance. While there are many projects using this workflow successfully (Wikipedia would probably be the most famous example), one such project directly related to audio archiving, namely the Gracenote MusicID (formerly called CDDDB) database [3]. Originally developed as a free application, MusicID has matured into a strong commercial utility for the cataloging of audio compact disks. MusicID and its free spinoff projects (e.g. freedb) can provide immense savings in time for any archive needing to ingest a substantial

number of audio CDs. Instead of having to enter track lists manually from the cover, the user just ingests the audio material and track data is downloaded from the internet. MusicID compatibility has become an indispensable requirement on most new audio archiving installations due to its saving potential. As a commercial service it now also provides a certain amount of editorial protection (unlike the free databases in which any user can overwrite the data input by another, often intentionally or inadvertently lowering the quality of the annotation) while still harnessing the power of the collective CD listener base worldwide.

Another source of cataloging data also belongs somewhat in this section even though it is not so much a collective effort but a centrally managed one-time process that drives the metadata availability: In the past, news audio was recorded from feed lines or field reporters. Today, most broadcasters receive the majority of their news audio digitally from agencies who make use of modern file formats like XML and MXF to not only transmit the audio itself but also a substantial amount of metadata, often not only involving keywords, person, date and location information but also complete news-reader transcripts ready for editing or on-air broadcast. In the future, it is to be expected that the exchange of fully cataloged audio via an appropriate file format (e.g. BWF, MXF, MPEG-7) will soon completely replace that of unstructured audio data. As a result, the core cataloging work will be done at the source once and receivers will only need to add some additional information mandated by their internal cataloging rules.

2.2.3. Speech to text transcription

Probably the most talked about application in research and the one that was given the most time and effort by developers is the transcription of speech to text. This field has made significant progress in the last twenty years and the situation presents itself as follows today:

- Speech recognition systems that require training on one specific speaker and a controlled environment have mostly peaked in their abilities for a few years. Depending on manufacturer, word recognition rates are in the range of 95 to 98 percent. These systems are nearly exclusively used in PC-based dictation systems since the conditions required for them to function are confined to studio work in the broadcasting context and the vast majority of text

spoken in this scenario is read from a script, thus invalidating any need for speech transcription.

- The second class of speech transcription systems is also requiring a relatively controlled, low-noise environment but does not require a specific speaker. Their recognition rate is approaching that of a speaker-trained system as long as the speakers have a reasonably accent-free language. Their primary use in an archiving scenario is for political speeches and debates that usually provide a reasonably noise-free background during the times the speaker actually says something. Recognition rates are not yet good enough to produce fully usable transcripts, combined with word lists and other criteria, they can however provide automated keyword extraction. These systems can also be used in conjunction with a transcript to match the transcript to the actual speech so that a user having found an area of interest in the transcript can immediately navigate to the appropriate audio section
- Unlike the other two classes, work on “general purpose” speech recognition systems that can cope with any speaker as well as background noise is still in a state where a commercial use is not viable. Word accuracy rarely exceeds 70 percent even in relatively controlled situations like that of an office in which someone is typing in the background. There are not yet any algorithms that can cope with multiple speakers at a time, even if one speaker is louder or closer than the others. Most likely, the desirable application of generalized speech to text transcription will not yet be commercially viable for several years to come.

In summary, speech to text systems have managed to recently achieve a stability that makes their use viable in some important newscasting situations. Especially parliamentary debates can now be indexed automatically with a reasonable accuracy and thus news programs can be supplied with searchable text quicker than in the past. Workflow integration of these systems is uncritical, so there is no real obstacle to their use. However, it should be noted that there is an alternative that provides the same searching functionality with higher accuracy at the expense of not having a full written text available. This alternative shall be discussed in the next section.

2.2.4. iFinder

iFinder is a novel approach to speech-to-text systems by the Fraunhofer Institute for Media Communication (IMK). [1] [2] Instead of the conventional strategy of making a one-pass transcription of speech and only then searching on the result, iFinder uses the phoneme level as its main retrieval source. The idea behind this is based on the following premises and observations:

- While there is always the possibility for misdetections on the phoneme level, the typical misdetection will not result in a random phoneme being returned but rather one that is similar to the intended one. This advantage is lost once the next step of transcribing the phonemes to standard text is undertaken. At this point, the misdetection either gets corrected by the algorithm or a completely wrong result is provided. Also, since nearly all successful speech-to-text systems use a dictionary to recognize valid words, acoustically small errors like the misdetection of a word boundary can result in completely wrong text being returned. Unknown words (e.g. names) will often not be transcribed at all, instead being forced to the "nearest word". By staying close to the original material, the iFinder algorithm avoids the cascading of the inevitable errors from the first processing step.
- There are very good algorithms for transforming text into speech or phonemes. Ambiguity rarely occurs when having a correctly spelled word and attempting to derive its pronunciation. The few words that actually have multiple pronunciations can either be evaluated from the context or fed to the phoneme comparison engine in both possible versions. Thus, iFinder replaces an inaccurate and unreliable process with one that is very exact.
- Fuzzy-match and relevance detection algorithms are just as easily applicable to phoneme sequences as they are to text. The number of phonemes in most human languages is not significantly different from the number of letters (European languages like English have between 30 and 50 distinguishable phonemes, a number well in the same order of magnitude as the 26 letters). The main difference is that phonemes contain significantly more information in a fuzzy match situation than letters. On the letter level the words "sight", "light", "night" and "might" all have the same distance from each other, namely one letter. On the phoneme level, n and m are very

closely related (and thus the easiest to possibly misinterpret when doing a speech-to-text analysis), l is already further away from them and the voiceless s bears no similarity to the other three letters. Thus, a relevance analysis values the most common misdetections close to the correct word and thus can include them in the search results.

By using these paradigms, iFinder achieves word recognition rates significantly beyond 98% in speaker-independent, untrained scenarios even with moderate background noise. It should however be said that this word recognition rate is not comparable to a word transcription rate since iFinder employed in this configuration does not actually produce a transcribed text. What it does however is to provide speech data in a representation that can be searched for using conventional, text-oriented algorithms with only minimal adaptations and thus achieve a search performance equivalent to a system that would actually transcribe text with the same accuracy. Also, the system is very easily integrated into a conventional database, ensuring a very easy transition from research into real-world applications and it – or systems with similar ideas – should be looked at very favorably by manufacturers of archive systems since they offer new functionality with minimal changes to the system.

3. NON-TEXTUAL TECHNIQUES

By their nature, most non-textual algorithms have both an annotation component and a retrieval component. I have grouped these in the section for non-textual retrieval. Before looking at those, there is however one algorithm based on generating and comparing text-like data from audio although, unlike true textual data, the derived data is not human-understandable.

3.1. Audio Fingerprinting

Audio fingerprinting is a method of analyzing audio and deriving from it a short, easily handled dataset that can then ideally uniquely identify the audio object in question. Two types of audio fingerprinting exist:

- Exact fingerprinting is based on commonly available hashing algorithms like MD5. It is not a technique unique to audio, it is rather usable to compare any group of files. With audio files being very large compared to text or image files, duplication detection in an archive would be prohibitive in processing effort involved. However, by generating and storing

an MD5 hash for each file, new files can quickly be checked to ensure the audio in question is not yet available in the archive. The main problem with MD5 and similar hashing algorithms is that a single bit change already completely changes the checksum value and thus any operation – even content-neutral operations like volume normalization, cutting off silence at the beginning or end, transcoding between formats – will invalidate the hash even though for human purposes, the audio is identical. Also, fuzzy matching is not possible on this kind of hash, preventing similarity detection.

- Heuristic fingerprints are based on the actual acoustic content of the audio and derive a measure from the actual usable content. For a single file, characteristics like formant frequencies, temporal distances between peak levels, envelope curves and similar criteria result in very distinctive and yet easily handled measurements. For sequences of files, the lengths of the individual files also provide good pre-sorting criteria, especially to identify CD albums. Audio fingerprints of this nature are ideal candidates for fuzzy matching technologies in order to detect close matches. This type of audio fingerprint may trigger false detections but provides extremely good identifications. It seems (although this is unconfirmed) that Gracenote MusicID uses a set of such fingerprinting algorithms to identify compact discs which renders the algorithm resistant to small changes introduced by new pressings (where silence lengths in tracks may be off and level normalization might vary between various issues of a CD). The author has personally experienced instances where a CD-R created from a vinyl record was correctly identified inside MusicID.

The practical usability of exact fingerprinting algorithms is minimal and restricted to situations where multiple storage of exactly identical items is to be prevented (such situations may occur in news-related environments where many editors work on the same raw material). Heuristic fingerprinting provides immense benefits when combined with collective annotation efforts and is thus rightfully already part of commercial solutions via the MusicID system.

3.2. Non-textual retrieval techniques

3.2.1. Query by Example

By a strict definition, a query by example cannot ever yield a first relevant result but it can, once a single relevant result is found, aid the user in finding other sounds that are relevant as well. Query by example answers the question of “I need a sound somewhat like this”, with “this” being a sound already retrieved from the archive. Depending on the application area, the algorithms used are quite different:

- Query by example on ambient sounds and noises is probably the most important application area. Media producers own large sound effects libraries from which they can pull and use sounds to match their needs. In many cases, a specific sound will be close to what they are looking for but they wonder if there is something similar that would fit better. A query by example then compares the sound by envelope, pitch variations, length and other criteria and selects those sounds that have an acoustic similarity. The FindSounds.com website [4] provides very good examples that also illustrate the criteria of what can make a sound actually be perceived as similar. Testing the database against an angry cat’s meow can find, amongst other cats meowing, such seemingly unrelated sounds as a race car (very similar envelope curve) and a key on a touch-tone phone (same principal frequencies), yet in both cases the acoustic similarity is immediately apparent.
- Query by example on spoken text would, if applied to the textual content, essentially be a word similarity check on the textual transcripts. Thus, it is not looking to be a viable application in the near future. What is however possible and can have useful applications is to attempt and analyze a voice speaking to find more text spoken by the same person. As speaker identification technology is far from perfect today, this does not promise any applications in the present or the near future, however in the middle to long term, speaker identification may allow sound users to query a database against a short sample of text from an unidentified speaker and retrieve a high-confidence return set of just a few possible names for the person and, as a result, providing access to more text by the same speaker.
- In music identification, query by example can detect musical pieces with similar melodies, similar rhythm

and similar instrumentation. Since these are detected with independent algorithms, a query by example on music can use rather different priorities in the quest to locate the matching pieces. Applications for this technology do not only lie in archival use to find songs matching an already found one in one or more criteria but also provides musicians with a potential way to detect plagiarism or protect themselves against any such suspicion by running their own piece against the database and listening to the closest matches.

Query by example technology is difficult to implement yet very easy and intuitive to use if the user knows what to expect from it. To stay with the example of the cat's meow above, many users would expect to receive a large collection of cat sounds as a result of a "sounds similar to this" query, but only a minority will actually have anything to do with felines. This contradiction evaporates once the user realizes that a similar sound is one with acoustically similar properties, not one that the high-level cognitive processes of the brain have classified into the same class of sound events. And if query by example actually did retrieve the sounds matching the naïve interpretation, it would invalidate its purpose since that result would just as easily be achieved by textually searching the annotation to select cat sounds from the database. The most viable commercial use of this technology in the short run is to aid users of large sound effects libraries in finding just the one effect they want. However, users need to be aware that this query technology will only yield sensible results when applied to large collections of audio since only close matches provide an actual benefit to the user. A more niche-oriented potential lies in the use of example querying on songs to deal with the plagiarism problem. A future application potential – which has not been realized so far – would be in the combination of the ideas of example querying and the iFinder concept to enable a system to locate identical or similar quotes based on an existing speech snippet.

3.2.2. Query by Humming

Query by Humming is technically a specialized application of example-based query. It is applicable only for the cataloging and retrieval of music. Query by humming can be implemented in two significantly different workflows:

- An easy to maintain but relatively unreliable query by humming implementation is based on a purely

melody-oriented implementation of query by example [5]. The system attempts to extract the main melody notes from all archived material, cataloging them by pitch and length. Then, the input is analyzed the same way (with a much lower error rate however since there is no accompaniment or secondary melody involved) and evaluation takes place. Problems with this setup include for one that it is very hard for a system to truly identify a melody note when there are multiple voices and instruments present and secondly that, even if the melody as such is identified correctly, there is no automatically detectable distinction between the key part of the melody (e.g. the beginning of the refrain) and any other melodic sequence. Combined, these two problems can and often do lead to hits being triggered by accidental matches of an input melody against a random sequence in a song. This implementation thus works well only for relatively simple songs with a clear lead voice and can benefit significantly from manual intervention in the annotation process.

- An implementation that is much more demanding in terms of annotation efforts but vastly superior in results relies on manual input and correction throughout the annotation phase. Supported by automated analysis, the annotator identifies key sequences in the melody, corrects misdetections (usually by selecting one out of multiple possible notes) and delivers a manually corrected or generated match set. For complex musical pieces, especially classical music and jazz, this is the only viable option. The matching rates and confidence of a search by humming are of course significantly superior to the other alternative, the search is also much faster and the database smaller (since only the key parts of each melody are stored and compared against) but the maintenance cost is very high.

Regardless of which of the two possibilities is used, the accuracy of a sung or hummed query always strongly depends on the talent of the person using the system. If they are able to maintain a good pitch and rhythm, the results are very promising, weak singers will often get poor results with no relevant matches. Fuzzy matching technology can eliminate some of these problems but also increases the risk of misdetection. One interesting approach for a database to be used by consumers was presented a few years ago on a trade show (unfortunately no reference to these seemingly promising results could be located by the author, thus ownership of that solution is unknown). In this

particular case, each melody was further reduced to a very simple fingerprint that allowed even an extremely weak singer to correctly reproduce the intended song (as far as the algorithm is concerned): Rhythm was completely ignored and pitch variation was quantized in only three flavors: "Same pitch", "Rising pitch" and "Falling pitch". With ten input notes, this algorithm only allows for $3^9 = 19683$ distinct fingerprints (the first note is neutral since it has no note to compare to, thus only nine contribute to the actual fingerprint), yet it has been found that the distribution of the key melodic elements of real musical works is very flat across this base – a selection of approximately 10000 popular works from multiple genres with a clearly identifiable, singable melodic theme would be unlikely to even have five pieces with identical fingerprints (even taking into account that some works need two fingerprints, e.g. the start of the verse and the start of the chorus). From the result set, the searcher can thus quickly, by prelistening the results, select the one work he was actually looking for. The author of this paper would be grateful for any new information on this promising technique which could easily be applied to systems targeted at providing consumers with music, especially if the owners of the works being cataloged would be willing to provide the melodic fingerprint information along with other metadata (they have the easiest time doing so). A retrieval technique similar to this could well provide an additional business impetus to music download endeavors by allowing consumers to locate and purchase a song that was just going around in their head.

4. CONCLUSION

Looking at the various techniques and algorithms [6], it becomes apparent that technology has produced a significant number of interesting products, but most of them are applicable only to specific types of sounds (if only by being restricted to one of the three main categories of music, spoken word and sound effect). These technologies provide a significant, often immediate benefit in better and/or faster cataloging and more precise and/or faster retrieval for the market niche they do apply to. In most other scenarios however, they are inapplicable. This discrepancy is the most striking reason why there is no widespread integration of most of these techniques into general-purpose audio archiving products. The main exception from this rule is the Gracenote / MusicID system that has found widespread acceptance among media users both on the consumer side and on the professional side as very few media

endeavors go completely without music CDs and the integration of MusicID is fully automatic, not modifying the search workflow at all and changing the annotation workflow only by possibly skipping the step of entering disc and track titles.

It is the author's opinion that especially professional systems can benefit from additional technologies even if those technologies apply only to a part of the audio holdings. Integrating a system like iFinder is likely to provide a current affairs editor with faster and more accurate results when attempting to locate older spoken material and especially in news and news-related programming, faster access means faster-to-air time at higher quality. Technology is at a point where developments have, either in the originally intended way or a completely different one, yielded results that can be integrated into an existing workflow with few changes. Vendors should consider their primary markets and offer some additional options in their products beyond simple searching for keywords. Customers should analyze their needs and decide on one or two tools that have the potential to directly improve their work and then ask vendors and integrators specifically for these selected features. It would be neither manageable nor desirable to equip an audio archive with every technology mentioned in this paper, but selective introduction of advanced technologies – be they the comparably simple text-based tools or the complex audio comparisons of an example-based query – can provide archive users with a quick return on investment without disrupting the existing workflow.

5. REFERENCES

- [1] <http://idw-online.de/pages/de/news60299> (in German)
- [2] http://www.imk.fraunhofer.de/sixcms/media.php/130/ifinder_eng.pdf
- [3] http://www.gracenote.com/gn_products/music_id.html
- [4] <http://FindSounds.com/>
- [5] <http://www.nuc.tu-berlin.de/qbh/>
- [6] <http://mirsystems.info/>