Spoken Content Metadata and MPEG-7

J.P.A. Charlesworth and P.N. Garner
Canon Research Centre Europe, 1 Occam Ct, Surrey Research Park, Guilford GU2 5YJ, England
+44 (0) 1483 448844
{jasonc,philg}@cre.canon.co.uk

ABSTRACT
The words spoken in an audio stream form an obvious descriptor essential to most audio-visual metadata standards. When derived using automatic speech recognition systems, the spoken content fits into neither low-level (representative) nor high-level (semantic) metadata categories. This results in difficulties in creating a representation that can support both interoperability between different extraction and application utilities while retaining robustness to the limitations of the extraction process. In this paper, we discuss the issues encountered in the design of the MPEG-7 spoken content descriptor and their applicability to other metadata standards.

Keywords
Spoken Content, MPEG-7, automatic speech recognition, spoken document retrieval, interoperability, robust retrieval.

1. INTRODUCTION
MPEG-7 is a metadata standard [1] for describing the content of multimedia documents, in particular those covered by the existing MPEG-1, 2 and 4 standards. Metadata standards are those that provide a framework enabling the description of the contents and meaning of a multimedia document rather than encoding the document itself. The metadata associated with a document may range in abstraction from low-level (representative) to high-level (semantic) data. Examples of low-level metadata are the dominant colour of an image or the Fourier power spectrum of audio, whilst examples of high-level metadata are a person's name or the emotional content of an image. This distinction in levels is characterized by differences in extraction, representation and specification. Low-level descriptors are typically unique (normative) algorithmic transformations of the multimedia document into a smaller parameter space. Such transformations may be expressed in mathematical terminology and performed automatically in software or hardware. The content of such low-level metadata is seldom readily interpretable by a human. In contrast, high-level descriptors, typically represent highly interpretive abstract human concepts for which there is no unique mapping. Such metadata must be specified in terms of words and cannot readily be created except through human intervention.

With the development of more powerful computers and extraction methods, the distinction between semantic and representative metadata is blurring and a middle level of metadata is emerging. Examples of such mid-level metadata are automatically derived spoken content of audio, topic identification in text and object identification in images. This metadata is characterized by being created by highly complex extraction tools attempting to resolve ambiguous many-to-many mappings. This results in imperfect extraction and the need for contextual cues to disambiguate the semantic component, e.g., in non-canonical English, the sounds of the words "picture" and "pitcher" may be identical but their meaning differs. Likewise, many different audio signals may represent a particular word. The meaning can only be disambiguated through topical or positional context. The extraction tools are extremely complex and not readily reproducible between annotations or between annotation and application due to the large number of free parameters. Were the metadata formed by a human annotator, it would be classified as high-level. However, automated extraction tools either through lack of contextual information or inability to utilize such information will be incapable of interpreting the data in the same manner as a human. Although mid-level metadata is evidently imperfect, the cost benefits of automated extraction ensure that its prevalence will increase.

It is essential that strategies for representing this data be developed. These must be robust to failures of annotation and support interoperability.

One of the most obvious and intuitive forms of mid-level metadata extractable from an audio-visual multimedia document is the spoken content, e.g., the actual words spoken. Whether required for direct searching of keywords and phrases or as input into further metadata extraction, such as topic identification, the spoken content forms an essential component of the audio-visual description. This content may be extracted at a number of levels from phonetic subword units (*phones*[1]) through syllables to words.

ACM Multimedia Workshop Marina Del Rey CA USA

[1] Throughout this paper, we use the 'phone' rather than 'phoneme' to refer to subword units. In practice, the difference between acoustic-data derived (phones) and linguistically derived (phonemes) does not affect our arguments. Indeed, most ASR phone sets are derived from phoneme sets.

In this paper we describe the development of the SpokenContent descriptor, an audio component of the developing MPEG-7 standard, and discuss the design considerations necessary to create a useable mid-level metadata descriptor. In Section 2, we sketch potential uses of spoken content metadata; these will be used to illustrate design decisions. In Section 3 we discuss the limitations of automatic speech recognition, (ASR), and the implication this has for interoperability and usability. In Section 4, we describe our SpokenContent descriptor and, in Section 5, we give experimental validation of this structure. In Section 6, we discuss the generality of such design considerations.

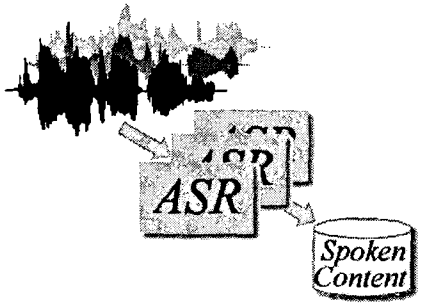
2. USES OF SPOKEN CONTENT

To illustrate the design considerations, consider two scenarios namely audio photographic captioning and video annotation.

The emerging digital photo standards [2] and, additionally, the convergence of digital photograph and video camera hardware, supports the annotation of images with compressed audio. Although it is unrealistic to assume that a user will create textual annotations for photographs at a later stage it is conceivable that a verbal annotation could be recorded at the time the photograph was taken. A typical personal photograph is of a person or object with the subject word forming the essential element of the annotation. We may thus expect photographic annotations to be short, all produced by a single speaker, of varying ambient noise and with a high percentage of people and place names[2]. Images would be downloaded to a server where the audio would be decoded using ASR trained to the user's speech. This would be used as the spoken content metadata attached to the image. The annotations would be used later by searching for key words either through spoken or textual queries.

Video annotation provides a different extreme. The annotation of, e.g., a film, will represent many speakers, potentially overlapping speech, possibly speaking more than one language. The audio will be long. The spoken content metadata may be derived in a studio using state-of-the-art systems. This metadata could be used for identifying portions of the video in which a particular subject was being discussed or where a particular interaction occurred between speakers. Such interaction may occur over a network as part of the selection process prior to downloading the multimedia document.

These means of interaction may be illustrated in figures 1 and 2.



[2] Typical photographs, along with textual and audio captions included as part of the MPEG-7 development set are available on www.cre.canon.co.uk/mpeg7/mebourne_photo_database.htm

Figure 1. The spoken content is derived from one or more audio streams containing speech. These are decoded using one or more speech recognition system to produce a single spoken content.

These scenarios, although superficially similar, make different demands on the accuracy of the automatic speech recognition system and consequently result in differing design considerations for the spoken content component of the MPEG-7 standard.

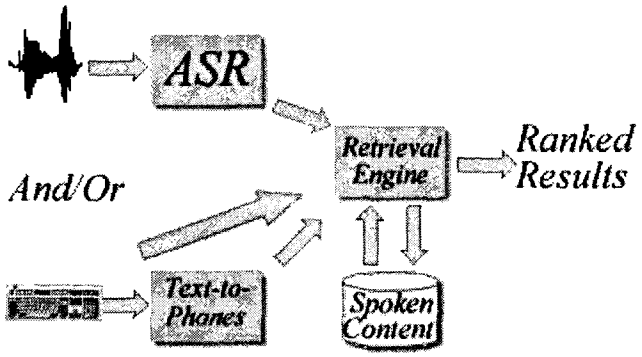


Figure 2. Retrieval may be performed using query-by-example or textual queries. Note that the ASR used for retrieval-by-example need not use the same engine or parameter set as that for annotation.

3. LIMITATIONS OF ASR

Except in the rare case of hand transcription, the spoken content will be the output of an automatic speech recognition system. Unlike most audio transformations, the output of an ASR engine must be considered a stochastic corruption of the idealised transformation (transcription). That is, we must consider the decoding of an audio signal into words to be a lossy and context dependent transformation. Even in ideal circumstances, it will not always be possible for the ASR system to extract the true spoken content [3]. In non-ideal, real-world or conversational speech, accuracy is limited by ambient noise, out-of-vocabulary words, ungrammatical construction and poor enunciation, leading to word recognition accuracy varying between 30 and 70% [4]. As such, in the design of a spoken content descriptor, especial attention must be paid to both the limitations of current ASR systems and the methods by which the metadata may be utilized for retrieval or other purposes.

From the premise that retrieval by keyword will be the most common use to which the spoken content will be put, we may identify two significant issues: extraction failures and extraction limitations.

Extraction failures. Internally, ASR systems store decodings in the form of lattices. These compactly represent a large number of hypotheses (see Figure 3) and may retain the correct decoding when the most probable decoding is in error. Recall may thus be improved by retaining some or all of the multiple hypotheses in the metadata. This is especially important for short audio captions, e.g., photographic annotations (see, e.g., the experimental results given in Section 5). For large audio documents, the value of multiple hypotheses (measured using precision rather than recall) is less clear [5], being masked by IR techniques.

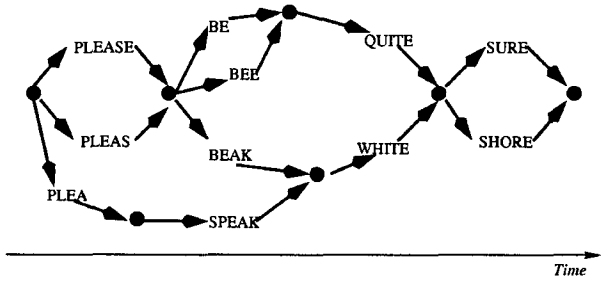


Figure 3. Hypothetical lattice representing the phrase "please be quite sure.

Extraction limitations. An ASR system typically employs a vocabulary of 20-60,000 words. Consequently, many nouns will be out-of-vocabulary. These form important discriminative retrieval terms, e.g., people or place names. Retaining the phonetic representation of the sound may facilitate retrieval of such terms through retrieval by example. This may be supported through a combined word and phone retrieval [6], e.g., utilizing the lattice illustrated in Figure 4.

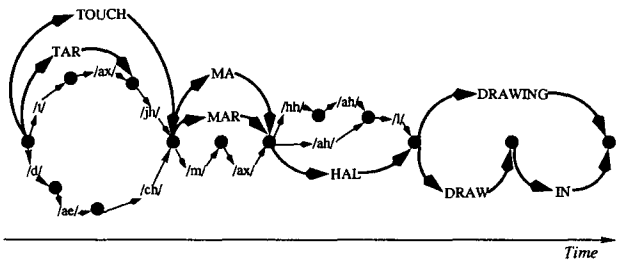


Figure 4. Hypothetical combined word and phone lattice structure for the phrase "...Taj Mahal drawing...". It is assumed that "Taj Mahal" is out-of-vocabulary.

The inclusion of the phonetic representation introduces interoperability issues discussed in Section 4.

It should be noted that an ideal transcription would require neither multiple hypotheses nor phonetic representation. This level of redundancy is a result of attempting to provide the applications with methods of circumventing the imperfections in the extraction utility.

4. MPEG-7 SPOKEN CONTENT DESCRIPTOR

In the previous Section, we described the failures of current ASR systems. We believe that to provide useable metadata we must represent the speech as a combined word and phoneme lattice, illustrated in Figure 4. Links between temporal nodes represent words or phones along with their likelihood. In practice, a single multimedia document may involve more than one spoken annotation, e.g., an audio annotated photographic library. Consequently, the spoken content will comprise multiple lattices with links attaching them to other metadata. Information common to all the lattices is stored in a separate header.

Although this structure ensures that the spoken content is retrievable, it addresses neither the issue of usability nor that of interoperability.

4.1 Usability

In itself, the lattices will not form an adequate level of metadata. An additional layer of metadata is required to permit valid interpretation and use of the spoken content lattices. For the example of video annotation, it may prove unrealistic to search the entire spoken content lattice representing a long film for a specific key word. Likewise, a simple textual representation of a word omits language information; is the word 'hat' an item of headwear in English or the verb 'to have' in German?

In the MPEG-7 SpokenContent, this explanatory set of metadata is stored in the header. Each speaker is associated with a language, a word lexicon, a phone lexicon and, optionally, word and phone indexes. As different applications may prefer different indexing schemes, the indexes provided are the simplest possible, i.e., word to location mapping and phone N-gram to location mapping. All other indexes may be constructed from these.

4.2 Interoperability

The issue of interoperability arises due to the complexity of the non-normative metadata extraction process; a state-of-the-art ASR system may use many millions of parameters to accurately model the statistics of a single language. We may not assume that the ASR system (or same parameter set) used for the metadata extraction is the same as that used for retrieval-by-example.

For the case of word matching or retrieval, the vocabulary ranges of the two may differ such that a query produced by one ASR engine (or equivalently entered by keyboard) could never be in the SpokenContent produced by the extraction utility. The only approach is to search the phonetic metadata of the SpokenContent. Like word decoding, ASR generated phonetic decodings are imperfect. However, the small number of phones, e.g., ~45 for English, means that statistics giving the confusability of the phones may incorporated in the SpokenContent as explanatory metadata. These confusion statistics permit statistical matching of phone strings from two, or more, different sources within a sound mathematical framework [7].

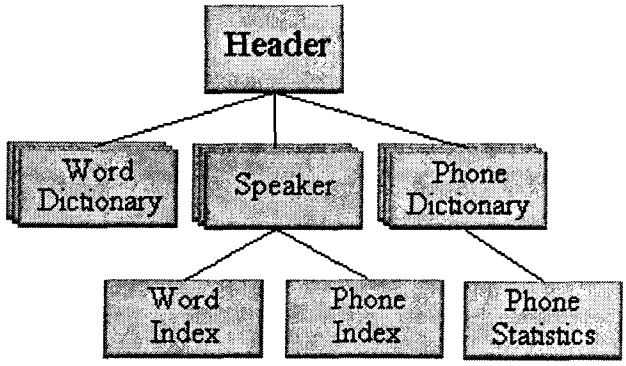


Figure 6. Schematic of the header component of the MPEG-7 spoken structure.

The inclusion of confusion statistics for our back-up, sub-word, representation is thus essential for interoperability between different extraction and application utilities. For this to be practical, a common phone set must be employed for each language. Although a language independent international phonetic language (IPA) exists it is preferable to use the SAMPA phoneset

derived from the IPA, this being more compact and maps better to the phone sets typically used in ASR systems.

In order to ensure both usability and interoperability we thus require the additional explanatory header metadata shown in Figure 6. For a full description of the MPEG-7 SpokenContent descriptor, see [8]. In typical SpokenContent descriptors, the header metadata is of comparable size to that of the lattice.

One benefit of the separation of header from lattice is that of incremental searching over networks, i.e., rather than transferring an entire annotation across the network, a partial search may be performed through downloading just those elements in a given language or by a particular speaker.

5. EXPERIMENTAL RESULTS

The fundamental requirement of any spoken content metadata is that it must store the spoken content in such a method that it may be quickly and reliably used at a later stage. The metric chosen to demonstrate this will depend on the application scenarios. Thus far, academic applications of spoken content have been largely dominated by the NIST funded spoken document retrieval thread of the TREC series of conferences [9]. In these, the number of query words is large and the documents long. MPEG-7, however, being designed for public usage will suffer from the problems illustrated in Section 2, e.g., short queries and (potentially) short annotations.

We have performed annotation and retrieval experiments for the two scenarios given. Fuller details are given in [10].

For the scenario of retrieval from the video soundtrack we found that average recall for a 2-word query rose from 0.57 for phones only and 0.67 for 1-best words only to 0.76 for a word lattice and 0.90 for combined word and phone lattice. The inclusion of phones in the metadata provides an essential back-up for cases where keywords are omitted in the annotation; in many cases the word-based query was unable to provide any match despite the query word being clearly audible in the sound track.

Likewise, in the photographic annotation scenario, the average recall for a 2-word query rose from 0.41 for 1-best words only and 0.43 for phones only to 0.51 for words and phones. The lower retrieval results here are indicative of the problems created by short annotations.

We have thus demonstrated that a multilevel approach to storing spoken content is essential for practical applications.

6. SUMMARY

We have demonstrated in this paper that the use of automatic speech recognition systems to extract the spoken content of an audio-visual multimedia document results in metadata midway between the semantic high-level metadata and the representative low-level metadata. Due to the imperfections of the extraction process, a complex multi-level representation must be employed to ensure that sufficient information is retained. This complexity, requires explanatory metadata to ensure interoperability between annotation and application utilities.

The general issues raised in the development of the MPEG-7 SpokenContent metadata descriptor are:

- *Multi-level representation.* Where the limitations of the extraction process mean there is a significant probability that the correct interpretation will not be retained, multiple interpretations need to be supported with the relative weights of the differing paths indicated. Nonetheless, to ensure retrieval where the decoding is outside the scope of the extraction utility, a lower-level representation needs additionally to be retained. We thus require a multi-level representation.
- *Usage Interoperability.* Mid-level metadata extraction methods utilize state-of-the-art systems undergoing constant advance. These tools are unlikely to be normative. We may not assume that two annotations were created both using the same extraction tool nor that the application utilizing the metadata employs the same extraction tool. Consequently, additional, explanatory metadata must be provided to ensure all annotations are treated equally. For the spoken content, phonetic confusion statistics are included to ensure, e.g., that the confusion probability of a canonical pronunciation is calculable independent of the extraction utility used in metadata creation.
- *Data fusion interoperability.* Whereas high-level metadata is often only amenable to binary matching, mid-level metadata yields a numerical match. To ensure that applications making use of multiple metadata descriptors can combine them in a consistent manner the metadata matching must be grounding in the language of statistics. In this way, e.g., word-based and phone-based retrieval results may be placed on same footing.

We believe that the issues raised in the development of the spoken content descriptor for MPEG-7 will be applicable in the development of other mid-level metadata descriptors.

7. REFERENCES

[1] See, e.g., www.mpeg-7.com

[2] See, e.g., www.digitalimaging.org

[3] For a comprehensive treatment of ASR techniques see Rabiner, L and B. Juang, *Fundamentals of Speech Recognition*, Wiley (1997).

[4] Johnson, S.E., et al., "Spoken document retrieval for TREC-7 at Cambridge University", Proc. 7th text retrieval conf., NIST special publication 500-242, p191 (1998).

[5] Siegler, M. et al. "Experiments in Spoken Document Retrieval at CMU", Proc. 7th text retrieval conf., NIST special publication 500-242, p319 (1998).

[6] Ng, K., "Information fusion for spoken document retrieval", Proc. ICASSP 4, p2405 (2000)

[7] Wechsler M, "Spoken document retrieval based on phoneme recognition" PhD thesis, Swiss federal institute of technology, Zurich (1998)

[8] Charlesworth, J.P.A., Garner P.N., Srinivasan S "Output of an of automatic speech recognition" ISO/IEC/JCC1/SC29/WG11 MPEG99/4458 (1999)

[9] The seventh Text REtrieval Conference, NIST special publication 500-242 (1998)

[10] Charlesworth, J.P.A., Garner P.N., Srinivasan S "Results of CE of automatic speech recognition" ISO/IEC/JCC1/SC29/WG11 MPEG99/5106 (1999)