# METADATA FOR INTEGRATING SPEECH DOCUMENTS IN A TEXT RETRIEVAL SYSTEM

Ulrike Glavitsch, Peter Schäuble, Martin Wechsler
Institut für Informationssysteme
Swiss Federal Institute of Technology (ETH)
CH-8092 Zürich (Switzerland)

## Abstract

We present an information retrieval system that simultaneously allows to search for text and speech documents. The retrieval system accepts vague queries and performs a best-match search to find those documents that are relevant to the query. The output of the retrieval system is a list of ranked documents where the documents on the top of the list satisfy best the user's information need. The relevance of the documents is estimated by means of metadata (document description vectors). The metadata is automatically generated and it is organized such that queries can be processed efficiently. We introduce a controlled indexing vocabulary for both speech and text documents. The size of the new indexing vocabulary is small (1000 features) compared with the sizes of indexing vocabularies of conventional text retrieval (10000 - 100000 features). We show that the retrieval effectiveness based on such a small indexing vocabulary is similar to the retrieval effectiveness of a Boolean retrieval system.

## 1 Introduction

We present a multimedia retrieval system that performs *content-based retrieval* from a document collection consisting of both text and speech documents. Whereas content-based retrieval is the retrieval strategy for many text retrieval systems, the idea of applying it to speech documents is very new [1], [5]. Our retrieval system accepts *vague queries* and adopts the *best-match retrieval* paradigm. A vague query asks for information about a given topic or a combination of topics. An example of a vague query may be "I am interested in documents about peace negotiations in the Far East". If the best-match retrieval paradigm is applied the documents are presented to the user in decreasing order of the degree by which they satisfy the search criteria.

The reason for using the best-match retrieval paradigm in our multimedia retrieval system is threefold. First, it can be shown mathematically that such a best-match retrieval approach maximizes the probability that a user finds the desired information [9]. Second, text retrieval systems having adopted a best-match retrieval strategy have achieved a high retrieval effectiveness in many retrieval experiments [6], [18]. Third, exact-match retrieval of non-textual objects is inappropriate because of recognition errors as shown in the following.

Database management systems traditionally support *exact-match retrieval* which is perfect for cases where, for instance, we want to retrieve the employees whose salaries are greater than a given threshold. In multimedia database management systems an exact-match retrieval may not be feasible because words cannot be identified uniquely in speech documents. Similarly, objects in images are not recognized perfectly as well. In the following, we will discuss in more detail why a perfect recognition of words in speech documents is not possible with state-of-the-art speech recognition technology. We will not discuss the recognition of objects in images since it is not relevant to this paper.

A speech document consisting of a sequence of continuously spoken words is a digitized waveform. The waveform is influenced by the gender of the speaker, its pitch, the loudness of speech, and the prosody. Current signal processing technology tries to extract features of the speech signal that are independent of the aforementioned influences. These features are encoded into coefficient vectors that are derived from equidistant intervals. Of course, not all of the influences mentioned above can be eliminated by the signal processing stage.

The word modelling stage identifies sequences of coefficient vectors as a single word. Words are predominantly modelled by hidden Markov models [7]. Hidden Markov models are stochastic models whose parameters are trained by means of special training algorithms. In addition, there exist well-evaluated recognition algorithms for both recognizing isolated words and continuous speech. The recognition algorithms return a probability for each detected isolated word and for a recognized sequence of continuously spoken words.

Extensions of these algorithms (wordspotting algorithms [10], [11], [19]) allow to compute the probability that a given word is spoken at a given position in a speech document. Current wordspotting algorithms only return positions where the probability is above a given threshold. If we attempt to search for all speech documents where a given word is spoken at least once, we will select all those documents where the wordspotter has detected at least one occurrence of the given word. We will add the probabilities of detection of each identified word in the same speech document resulting in a global probability of detection for each document. We then rank the documents in decreasing order of these global probabilities. Hence, the best-match retrieval paradigm is also important for multimedia databases because current speech and image recognition systems are not able to recognize words and images perfectly.

In what follows, we outline the retrieval method we are using in our text and speech retrieval system. The metadata consists of description vectors where for every document $d_j$ there exists a corresponding document description vector $\vec{d_j}$. A component of the document description vector represents the weight for a particular indexing feature. In the case of conventional text retrieval, the indexing features are the terms used to describe the content of a document. In our case, the indexing features consist of subword units as described in the next section. The generation of metadata is called the *indexing method*.

When a user submits a query $q$ to the retrieval system, the query description vector $\vec{q}$ is computed in the first step. The query description vectors are most often derived from queries in the same way as the document description vectors are derived from documents. To estimate the relevance of a document with respect to the query, a *retrieval function* $\rho$ is applied to the document description vector as well as to the query description vector. The output of the retrieval function is called *retrieval status value*, $RSV(q, d_j) = \rho(\vec{q}, \vec{d_j})$. A retrieval function which was shown to perform well in many cases is the cosine measure [8]:

$$\rho(\vec{q}, \vec{d_j}) := \frac{\vec{q}^T \vec{d_j}}{\sqrt{\vec{q}^T \vec{q}} \sqrt{\vec{d_j}^T \vec{d_j}}}$$

The RSV is high if the angle between the document description vector and the query description vector is small and the RSV is low if the angle between the two vectors is large. A small angle between the document description vector and the query description vector means that the document and the query have a lot of indexing features in common. The retrieval system presents a list of ranked documents to the user in accordance with the probability ranking principle as mentioned above.

The main contribution of this paper is the description of the metadata specifically designed for speech documents. We will show that the presented type of metadata can also be used for texts such that the proposed multimedia retrieval system is capable to simultaneously retrieve text and speech documents.

The paper is structured as follows. In Section 2, we present the metadata for speech documents. In Section 3, we elaborate on the metadata organization and the processing of queries. In Section 4, we assess the performance of the metadata used in the retrieval system. Finally, we draw some conclusions in Section 5.

## 2 Generating Metadata for Speech Documents

In this section, we show how metadata for speech documents is created. We will show that the metadata for speech documents are document description vectors similar to the metadata generated for texts. This form of metadata has the advantage that a retrieval function of conventional text retrieval can be applied and is expected to perform well. However, we defined a new indexing vocabulary that is suited to describe a collection of speech documents. We will see later that the same indexing vocabulary is also appropriate to index texts.

Indexing vocabularies of conventional text retrieval are not appropriate for the description of speech documents. Indexing vocabularies to describe texts are usually derived from the documents in a given collection during the process of generating the metadata, i.e., the indexing terms and the effective number of indexing features are not known in advance. If we are to describe speech documents, a speech recognition component is used to detect occurrences of indexing features in the speech documents. In this case, the indexing vocabulary must be *controlled*, i.e., the indexing features are fixed, since the speech recognition component requires speech models prepared for each indexing feature. The preparation of speech models is expensive since sufficient training data for each speech model is needed. The number of detected occurrences of each indexing features in a speech documents is a major parameter in the computation of the metadata of the speech document.

Both the speech recognition component and the retrieval process impose restrictions on the selection of appropriate indexing features. We formulated the requirements for an indexing vocabulary for speech documents as follows:

- The indexing vocabulary must be small, since the (manual) preparation of training data is expensive.
- The document frequency $df(\varphi_i)$ of an indexing feature $\varphi_i$, i.e., the number of documents in which $\varphi_i$ occurs, must be below an upper bound $df_{max}$. A feature with a very high document frequency occurs in almost all the documents and it is therefore less likely to be a suitable indexing feature.
- At the same time, the document frequency $df(\varphi_i)$ of an indexing feature $\varphi_i$ must be above a lower bound $df_{min}$ in order to ensure that there exists sufficiently many training samples.

Among the possible candidates for indexing features there are phonemes, biphones, words and phrases. Words seem to be too large a unit since the resulting vocabulary size would be too large. Phonemes, however, are too small. They occur too frequently in all the speech documents, hence they are not a good unit for indexing. Thus, good indexing features seem to lie between words and phonemes. We selected three special types of subword units to be used as indexing features: VCV-, CV-, and VC-features. The letter V stands for a maximum sequence of vowels and C for a maximum sequence of consonants within the same word. CV-features and VC-features occur at the boundaries of a word. A CV-feature only occurs at the beginning of a word that starts with a consonant whereas a VC-feature only occurs at the end of a word that ends with a consonant.

Our choice of indexing features was motivated by Teufel's work on trigrams [16]. Since vowels and consonants are defined for text and speech, the presented features can be detected in both media types. For example, all possible VCV-, CV-, and VC-features of the written word INTERNATIONAL are INTE, ERNA, ATIO, IONA, and AL. The features of the spoken word INTERNATIONAL are the speech units corresponding to INTE, ERNA, ATIO, IONA, and AL.

In the following, we present an algorithm to compute an indexing vocabulary for a particular domain:

- INPUT
  1. speech documents (i.e., audio recordings)
  2. text documents of the same domain
- OUTPUT
  1. 1000 VCV-, CV-, and VC-features for indexing
  2. pronunciations of these features to train the speech models
- ALGORITHM
  1. Parse the text documents to determine all VCV-, CV- and VC-features occurring in the texts.
  2. Determine the indexing vocabulary by selecting those 1000 VCV-, CV-, and VC-features $\varphi_i$ with document frequencies $df(\varphi_i)$ between $df_{min}$ and $df_{max}$ and with a small probability of being confused with any other of the selected features. Pairs of VCV-, CV, or CV-features are likely to be confused if they are similarly pronounced.
  3. Extract different pronunciations of each indexing features from the speech documents by segmenting manually the speech documents.

As soon as an indexing vocabulary for speech documents has been defined and as soon as the speech models for each indexing feature are trained, the speech recognition component can start by detecting occurrences of each indexing feature in the speech documents. The number of identified occurrences of each indexing feature $\varphi_i$ in a speech document $d_j$ is called the feature frequency $ff(\varphi_i, d_j)$. The document description vector

$$\vec{d_j} = (a_{0,j}, \ldots, a_{m-1,j})$$

consists of $m$ weights $a_{i,j}$ where $m$ denotes the number of indexing features. Each component of the metadata $\vec{d_j}$ represents a weight $a_{i,j}$ for the indexing feature $\varphi_i$. The weight for an indexing feature should be high if the feature characterizes the document well and it should be low if the feature is not important for that document. A weighting scheme fulfilling this requirement is the following [13].

$$a_{i,j} = ff(\varphi_i, d_j) * idf(\varphi_i)$$

The inverse document frequency $idf(\varphi_i)$ of an indexing feature $\varphi_i$ is an estimate of the specificity of a feature. The inverse document frequency is high if the feature occurs in a small number of documents and it is low if the feature is contained in almost all the documents. The inverse document frequency is defined by:

$$idf(\varphi_i) := \log\left(\frac{n+1}{df(\varphi_i)+1}\right)$$

where $n$ denotes the number of documents in the collection. The document frequency $df(\varphi_i)$ is the number of documents containing $\varphi_i$. It is formally defined as

$$df(\varphi_i) := |\{d_j \in D | ff(\varphi_i, d_j) > 0\}|$$

where the symbol $D$ denotes the set of all documents.

As mentioned above, the indexing features can be identified in both text and speech documents. Occurrences of VCV-, VC-, and CV-features are identified in texts by means of simple pattern matching. Thus, the same type of description vector is generated for documents of both media types. As a consequence, the retrieval system can simultaneously retrieve text and speech documents.

To summarize, the metadata for a collection of text and speech documents is generated as follows. First of all, an indexing vocabulary has to be computed according to the algorithm presented above. Secondly, we train the speech models for each indexing feature. Then, the speech recognition component detects occurrences of all the indexing features in the speech documents and computes the resulting feature frequencies. In case of text documents, the frequencies of indexing features in documents are easily determined by parsing the texts. Finally, we generate a document description vector for each document. The components of the document description vectors are computed according to a conventional weighting scheme.

## 3 Metadata Organization and Query Processing

The problem of evaluating best-match queries is the following. Given are the query description vector $\vec{q}$, the document descriptions $\vec{d}_0, \ldots, \vec{d}_{n-1}$, the retrieval function $\rho$, and a number $k$, find those $k$ documents $d_j$ having the highest function values $\rho(\vec{q}, \vec{d}_j)$. Because the response time should be less than one second, a naive comparison of the query description with every document description is inadequate when the document collection contains a few thousand documents or more. The standard approach to the evaluation of best-match queries takes advantage of

1. the fact that most documents have no indexing features in common with any query,
2. the fact that the retrieval status value $RSV(q, d_j)$ is minimum if the query $q$ and the document $d_j$ do not have any indexing features in common, and
3. possibly certain stopping conditions considering the maximum similarity that the query can have with documents not yet examined [2].

In the simplest form, the metadata is organized as an *inverted file,* i.e., as an array of lists of postings. The $i^{th}$ entry in such an array corresponds to the indexing feature $\varphi_i$ and this entry contains the following list of postings.

$$\{(j, a_{ij}) \mid a_{i,j} \neq 0\}$$

Given such an inverted file, the query processing looks as follows.

**FOR EACH** query feature $\varphi_i$ **DO**
    lookup the corresponding list of postings;
    **FOR EACH** posting $(j, a_{ij})$ in this list **DO**
        increment $RSV(q, d_j)$ by $a_{ij} b_i$;
    **END**;
**END**;
divide positive $RSV(q, d_j)$'s by $\sqrt{\vec{q}^T \vec{q}. \vec{d_j}^T \vec{d_j}}$;
sort the RSV's and return the top $k$ documents;

Conventional database systems do not support well the represention of the posting lists because they assume that the accesses are equally distributed which is absolutely not the case. In [3], a sophisticated buffering scheme is proposed which takes into account the access distribution.

In the case of frequent updates, inverted lists are usually inadequate because the search for postings in long lists is time consuming. For this case, an alternative access structure is proposed which is based on signatures and non-inverted document descriptions [8]. In [17], an approach is presented for the special case where updates mainly consist of insertions and not of modifications nor deletions. Considering this variety of different access structures, it seems that extensible systems will play an important role in supporting best-match retrieval [12]. See also descriptions of the "document blade" of the commercial product Illustra which is based on the Berkeley University Postgres DBMS.

## 4 Retrieval Effectiveness Affected by Recognition Errors

In Section 2, we presented how the metadata for speech documents is generated. However, the computation of the metadata is error-prone since the feature frequencies computed by the speech recognition component are not likely to be correct. The speech recognition component may not detect an indexing feature or it may locate one at a position where it is not present. These two types of error are called misses and false alarms. In this section, we show how much an error-prone recognition of indexing features affects the retrieval effectiveness, i.e., the performance of the metadata for retrieving both text and speech documents.

Before we computed the effectiveness for retrieving speech documents we were interested in the retrieval accuracy in the searching for texts. This case is equivalent to the retrieval of speech documents where the indexing features are recognized perfectly. In [5], we were able to show that the retrieval effectiveness of the retrieval method for texts is comparable to a conventional retrieval method even though an extremely small indexing vocabulary is used.

The retrieval effectiveness of retrieving speech documents is computed as follows. We simulated the process of generating metadata for speech documents on standard information retrieval text collections. The simulations were carried out on the CRANFIELD collection. We computed an indexing vocabulary according to the algorithm presented in Section 2. We incorporated two types of recognition errors into the simulations: misses and false alarms. A more complete description of the simulation of recognition errors can be found in [14]. The retrieval effectiveness is measured in terms of the mean precision [13]. In Table 1, we present the mean precisions for different detection rates and false alarms per keyword per hour (fa/kw/hr) achieved on the CRANFIELD text collection. The detection rate represents the percentage of correctly recognized indexing features whereas the number of false alarms per keyword per hour is self-explanatory.

The results show that a 80% or 90% detection rate in the absence of false alarms has only small effects on the retrieval effectiveness. This is in accordance with results achieved on OCR output [15]. With a growing number of false alarms per keyword per hour the retrieval effectiveness decreases significantly. However, at an operating point of 50% detection rate and 110 fa/kw/hr we achieve a retrieval effectiveness of boolean retrieval [4]. This is not a very exciting retrieval effectiveness but it is one where users still find useful information.

Current speech recognition technology is far from identifying speech units as small as the proposed indexing features reliably. As a result, the computed (noisy) feature frequencies differ largely from the correct feature frequencies. The question arises whether it is possible to correct the noisy feature frequencies such that a better retrieval effectiveness can be achieved. Based on the observation that the false alarms are more or less equally distributed over all the documents, we developed an analytical model to compute the expected noisy feature frequency of a feature in a document given the correct feature frequency, the detection rate and the false alarms per keyword per hour. In a first step, we calculate the probability that the noisy feature frequency $noisy\_ff_{ij}$ of a feature $\varphi_i$ in the document $d_j$ equals $z$ given that $ff_{ij}$ is the correct feature frequency:

$$P(noisy\_ff_{ij} = z | ff_{ij}) \quad = \qquad (1)$$

$$\sum_{x0}^{x1} A(dr, ff_{ij}, x) A(fr, l_j - ff_{ij}, max(0, z - x))$$

where

$$\begin{aligned} x0 &:= max(0, z - l_j + ff_{ij}) \\ x1 &:= min(z, ff_{ij}) \\ A(p,u,v) &:= \begin{pmatrix} u \\ v \end{pmatrix} p^v (1-p)^{u-v} \end{aligned}$$

denotes the binomial distribution, $dr$ denotes the detection rate, $fr$ is the probability that a false alarm occurs per average duration of a feature (false alarm rate), and $l_j$ refers to the length of the document $d_j$ measured in average duration of a feature. By means of the probabilities (1) we can compute the expectation value of the noisy feature frequency given that the correct feature frequency has a certain value.

A table containing quadruples of expected noisy feature frequency, correct feature frequency, detection rate and false alarms per keyword per hour is built. The noisy feature frequencies encountered in the process of metadata generation can then be corrected by looking up the corresponding correct feature frequencies in the table. Whether such a compensation for recognition errors improves the retrieval effectiveness or not is subject to current investigation.

## 5 Conclusions

We have shown that integrating speech documents in an information retrieval system is feasible. For the retrieval of speech documents the development of a new controlled indexing vocabulary was necessary in order to generate the required metadata. The size of the indexing vocabulary is small (1000 features) compared with vocabulary sizes of conventional text retrieval (10000 - 100000 features). Yet, the retrieval effectiveness of searching speech documents is acceptable even if recognition errors occur in the process of metadata generation.

Future work will concentrate on improving the recognition performance of the speech recognizer and the incorporation of relevance feedback into the speech retrieval system. Relevance feedback is an effective

CRANFIELD: average precision of the reference method: 0.408 (100%)

| $dr, fa$ | 0 | 10 | 20 | 50 | 80 | 110 |
|---|---|---|---|---|---|---|
| 90% | 0.330 (81%) | 0.315 (77%) | 0.304 (75%) | 0.288 (71%) | 0.279 (68%) | 0.264 (65%) |
| 80% | 0.324 (79%) | 0.299 (73%) | 0.291 (71%) | 0.276 (68%) | 0.265 (65%) | 0.253 (62%) |
| 70% | 0.305 (75%) | 0.283 (69%) | 0.267 (65%) | 0.265 (65%) | 0.242 (59%) | 0.249 (61%) |
| 60% | 0.297 (73%) | 0.253 (62%) | 0.234 (57%) | 0.245 (60%) | 0.249 (61%) | 0.224 (55%) |
| 50% | 0.277 (68%) | 0.236 (58%) | 0.224 (55%) | 0.226 (55%) | 0.211 (52%) | 0.206 (50%) |
| 40% | 0.259 (63%) | 0.216 (53%) | 0.191 (47%) | 0.185 (45%) | 0.191 (47%) | 0.175 (43%) |

Table 1: Average precision values for detection rates within the range of 40% and 90% and false alarms per indexing feature (key word) per hour within the range of 0 and 110. The numbers in brackets represent the percentage of the average precision of the reference method, i.e., a standard text retrieval method with a large indexing vocabulary.

technique to derive automatically a new query from the old query and from the user's relevance judgements. The new query is likely to retrieve more relevant documents.

## References


[1] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Sparck Jones, and S. J. Young. Video Mail Retrieval by Voice: An Overview of the Cambridge/Olivetti Retrieval System. In *Multimedia Data Base Management Systems' Workshop at the 2nd ACM International Conference on Multimedia*, 1994.

[2] C. Buckley and A. F. Lewit. Optimization of Inverted Vector Searches. In *ACM SIGIR Conference on R&D in Information Retri4 11:51 eval*, pages 97–110, 1985.

[3] D. Cutting and J. Pedersen. Optimization for Dynamic Index Maintenance. In *ACM SIGIR Conference on R&D in Information Retri4 11:51 eval*, pages 405–411, 1990.

[4] E. A. Fox and M. B. Koll. Practical Enhanced Boolean Retrieval: Experiences with the SMART and SIRE Systems. *Information Processing & Management*, 24(3):257–267, 1988.

[5] U. Glavitsch and P. Schäuble. A System for Retrieving Speech Documents. In N. Belkin, P. Ingwersen, and A. M. Pejtersen, editors, *ACM SIGIR Conference on R&D in Information Retri4 11:51 eval*, pages 168–176, 1992.

[6] W. Hersh. OSHUMED: An Interactive Retrieval Evaluation and Large Test Collection for Research. In *ACM SIGIR Conference on R&D in Information Retri4 11:51 eval*, pages 192–201, 1994.

[7] X. D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition*. Edinburgh University Press, Edinburgh, 1990.

[8] D. Knaus and P. Schäuble. Effective and Efficient Retrieval from Large and Dynamic Document Collections. In *TREC-2 Proceedings*, pages 163–170, 1993.

[9] S. E. Robertson. The Probability Ranking Principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[10] J. R. Rohlicek, W. Russel, S. Roukos, and H. Gish. Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting. In *ICASSP*, pages 627–630, 1989.

[11] R. C. Rose and D. B. Paul. A Hidden Markov Model Based Keyword Recognition System. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 129–132, 1990.

[12] P. Schäuble. SPIDER: A Multiuser Information Retrieval System for Semistructured and Dynamic Data. In *ACM SIGIR Conference on R&D in Information Retri4 11:51 eval*, pages 318–327, 1993.

[13] P. Schäuble. Multimedia Information Retrieval. Tutorial held at the ICMCS'94 conference, Boston, May 1994.

[14] P. Schäuble and U. Glavitsch. Assessing the Retrieval Effectiveness of a Speech Retrieval System by Simulating Recognition Errors. In *ARPA Workshop on Human Language Technology (HLT'94)*, 1994.

[15] Taghva, J. Borsack, and A. Condit. Results of Applying Probabilistic IR to OCR Text. In *ACM SIGIR Conference on R&D in Information Retri4 11:51 eval*, pages 202–211, 1994.

[16] B. Teufel. *Informationsspuren zum numerischen und graphischen Vergleich von reduzierten natürlichsprachlichen Texten.* PhD thesis, Swiss Federal Institute of Technology, 1989. VdF-Verlag, Zürich.

[17] A. Tomasic, Hector Garcia-Molina, and Kurt Shoens. Incremental Updates of Inverted Lists for Text Document Retrieval. In *SIGMOD*, pages 8–17, 1994.

[18] H. Turtle. Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance. In *ACM SIGIR Conference on R&D in Information Retri4 11:51 eval*, pages 212–220, 1994.

[19] L. D. Wilcox and M. A. Bush. HMM-Based Wordspotting for Voice Editing and Indexing. In *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 25–28, 1991.