PHONE-BASED SPOKEN DOCUMENT RETRIEVAL IN CONFORMANCE WITH THE MPEG-7 STANDARD

NICOLAS MOREAU, HYOUNG GOOK KIM, AND THOMAS SIKORA

Communication Systems Group, Technical University of Berlin, Germany
[moreau,kim,sikora]@nue.tu-berlin.de

This paper presents a phone-based approach of spoken document retrieval, developed in the framework of the emerging MPEG-7 standard. The audio part of MPEG-7 encloses a *SpokenContent* tool that provides a standardized description of the content of spoken documents. In the context of MPEG-7, we propose an indexing and retrieval method that uses phonetic information only and a vector space IR model. Experiments are conducted on a database of German spoken documents, with 10 city name queries. Two phone-based retrieval approaches are presented and combined. The first one is based on the combination of phone *N*-grams of different lengths used as indexing terms. The other consists of expanding the document representation by means of phone confusion probabilities.

INTRODUCTION

The profusion of audio-visual documents that are today available requires the design of indexing and retrieval methods able to exploit different sources of information (image, video, audio, text, etc.). In particular, many audio-visual documents contain speech signals enclosing a lot of useful information allowing to index and retrieve the documents they belong to.

A first way to exploit the spoken information is to let a human operator listen to it and transcribe it into textual information. In real word applications however, the hand indexing of spoken audio material is impracticable, owing to the huge volume of most databases. An alternative is the automation of the transcription process by means of an automatic speech recognition (ASR) system.

Each spoken document is in this case indexed by the recognition hypotheses delivered by a speech recogniser. Because the extracted spoken content can be output in different ways and formats, it poses problems of compatibility for systems which need to access and exchange this information. A new multimedia standard emanating from the Moving Picture Experts Group [1] offers a solution to that problem. This standard, called "MPEG-7" [2],[3] (or "Multimedia Content Description Interface"), aims at providing a unified way of describing the content of any kind of multimedia documents. In particular, the audio part of MPEG-7 [4],[5] contains a SpokenContent Description Scheme (DS) [6] consisting of a standardized description of the information extracted by ASR systems from spoken data.

This study presents a system to index and retrieve German spoken documents in conformance with the MPEG-7 specifications. We developed an indexing system based on the extraction of phonetic information. The use of sub-word (phones in our case) rather than words as basic indexing units makes the system vocabulary independent [7],[8],[9],[10].

However, phone recognition systems have to contend with high error rates. Classical text retrieval methods are not relevant in that case. More specific spoken document retrieval (SDR) methods are required. In this study, we consider two different ways of coping with the problem of phone recognition inaccuracy.

The first one consists of indexing the spoken documents with MPEG-7 phone lattices rather than 1-best transcriptions. This allows to take several competing phone hypotheses into account.

The other one takes advantage of some ASR specific information, the phone confusion statistics [7],[8],[10],[11], to compensate for the recognition errors. The SDR system presented here exploits the phone confusion matrix enclosed in the MPEG-7 SpokenContent DS for expanding the phonetic representation of documents.

The paper is structured as follows.

Section 1 introduces briefly the MPEG-7 SpokenContent description. We then discuss the benefits and drawbacks of phone-based indexing. Finally the features of our phone recognition system are detailed.

In section 2, we present the overall structure of a SDR system. We then describe different retrieval strategies based on the extraction of phone *N*-grams.

Experiments were performed on a database of German spoken documents in order to evaluate the retrieval strategies introduced before. The experimental results are reported and commented in section 3.

Finally, conclusions and several perspectives for future investigations are exposed in section 4.

1 SPOKEN CONTENT INDEXING

This part first gives a short insight into the MPEG-7 spoken content description. The indexing system used in this study, based on phonetic information in conformance with the MPEG-7 SpokenContent standard, is then detailed.

1.1 MPEG-7 SpokenContent Description

Most of nowadays ASR systems are based on the Hidden Markov Model (HMM) paradigm [12]. The HMMs can model any kind of speech units (words, phones, etc.) allowing to design systems with diverse degrees of complexity. In any case the idea is basically the same. The input data is matched against a series of word or sub-word HMMs by means of a Viterbi algorithm, constrained by a language model (LM). The sequence of models yielding the best likelihood score gives the best transcription hypothesis.

Actually, a speech recogniser can output the recognized sequence in several different ways. The delivering of a single recognition hypothesis is enough for the most basic systems (isolated or connected word recognition). But when the recognition systems are more complex, the most probable transcription usually contains several errors. In that case, it is necessary to deliver a series of alternative recognition hypotheses on which further post-processing operations can be performed. This information can be represented in two ways:

- a *n-best list*, where the *n* most probable transcriptions are ranked according to their scores.

- a *lattice*, consisting of an oriented graph whose paths represent different possible transcriptions. Each node in the graph represents a time point between the beginning and the end of the speech signal. A link between two nodes corresponds to a recognition hypothesis (e.g. a word).

A lattice can be seen as a reduced representation of the initial search space. It offers a compact description of the transcription alternatives and can be post-processed in many different ways.

Basically, the MPEG-7 SpokenContent tool defines a standardized description of the lattices delivered by a recogniser [6]. The Figure 1 gives an illustration of what an MPEG-7 SpokenContent description of the speech input "*Film on Berlin*" can be.

The standard defines two types of lattice links: word type and phone type. A MPEG-7 lattice can thus be a word-only graph, a phone-only graph or combine word and phone hypotheses in the same graph as depicted in the example of Figure 1. Each link is assigned a label (a word or a phone), a probability derived from the language model, and the acoustic score delivered by the ASR system for the corresponding hypothesis.
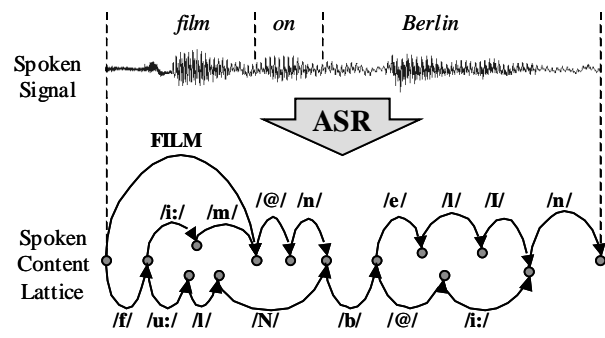

Figure 1: MPEG-7 SpokenContent description of an input spoken signal "*Film on Berlin*".

A MPEG-7 SpokenContent description may also contain some additional information that can be useful for filtering, indexing or retrieval applications:

- a word lexicon (if words are used).
- a phone lexicon (if phones are used).
- phone confusion information: a confusion matrix (with statistics on substitutions) plus statistics on insertions and deletions.
- diverse information about the speakers, the spoken language, the ASR system, etc.

As explained below, this study will only deal with phone graphs and the associated metadata: the phone lexicon and the phone confusion information.

1.2 Phone-Based Indexing

Many SDR systems proposed in the past consisted in linking a word recognition engine with a traditional text retrieval system [13]. In that case, the vocabulary has to be known beforehand, which precludes the handling of out-of-vocabulary (OOV) words. Essentially, the word-based approach has two drawbacks:

- The static nature and limited size of the recognition vocabulary (that directly restricts the indexing and query vocabulary).
- The derivation of complex stochastic language models (LMs) is necessary for reasonable quality LVCSR systems (large vocabulary continuous speech recognition), requiring huge amounts of training data.

In the recent years, other approaches have considered the indexation of spoken documents with sub-lexical units instead of word hypotheses [7],[8],[9],[10]. In that case, a limited amount of sub-word models is necessary, allowing to index any sentence (for a given language) with sub-lexical indexing terms, such as phones [8], phonemes [7],[10] or syllables [9].

In this study, we will only use phone graphs. Our indexing system does not require any *a priori* word lexicon. The use of phones as basic indexing terms restrains the size of the indexing lexicon to a few dozens of units.

However, phone recognition systems have a major drawback. They have to cope with high error rates,

typically comprised between 30% and 40%. In our case, the challenge is to exploit efficiently the MPEG-7 SpokenContent description in order to compensate for these high error rates.

1.3 Phone Recognizer

In the following, we describe the phone recogniser that was used for extracting the MPEG-7 phone lattices in the indexing system. It was employed for conducting the experiments reported in section 3.

1.3.1 Acoustic Models

The language considered in this study is German. We used a set of 42 phone symbols derived from the 46 German phones of the SAMPA alphabet [14].

Each phone, along with a silence unit, is modelled by a context independent HMM having between 2 and 4 states, depending on the phone. The observation functions are multi-gaussians with 128 modes per state and diagonal covariance matrices. We used 39-dimensional observation vectors (12 mel-frequency cepstral coefficients, the log energy, plus their first and second derivatives).

The HTK toolkit [15] was used for training the HMMs on the German "Verbmobil I" (VM I) corpus [16]. It is a large speech database consisting of spontaneous (non-prompted) speech from many different speakers and environments.

1.3.2 Phone Loop

The recogniser used for indexing performs phone recognition without any lexical constraints. The 43 context independent Markov models are looped, according to a bigram language model (LM). The LM has been derived from the transcriptions of the whole Vermobil II (VM II) corpus.

Given a spoken input, our ASR system produces an output phone lattice containing several hypothesized phonetic transcriptions. In order to reduce the set of indexing symbols, we mapped our 42 SAMPA phones to 32 German "phonemes" as proposed by [10], thus avoiding the distinction between very similar sounds (e.g. phones [a:] and [a] are merged to form a single phoneme class /a/). For more convenience, we will continue in the following to use the term "phone" instead of "phoneme".

The recogniser has been tested on the 14th volume of the German VM I corpus (VM14.1) which had not been used for training. This provided the phone confusion probabilities. The resulting confusion matrix is enclosed, along with the phone lexicon and the phone lattices, in the MPEG-7 SpokenContent descriptions.

2 RETRIEVAL

After giving an overall description of a SDR system, we present the different retrieval strategies that will be tested in the next section.

2.1 Spoken Document Retrieval

Once indexed, the documents can be processed by a spoken document retrieval (SDR) algorithm.

An example of indexing and retrieval system based on the MPEG-7 SpokenContent tool is schematically depicted in Figure 2. During the indexing phase, each spoken part of the audio-visual (AV) documents in the database is processed by an ASR system. When documents contain other types of sounds (e.g. music in the audio stream of a radio broadcast), a preliminary audio segmentation step is necessary to isolate the spoken parts.
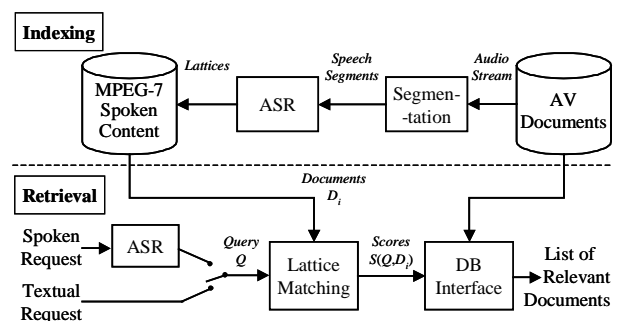

Figure 2: Spoken Document Indexing and Retrieval.

Once indexed, the audio-visual database can be processed by an information retrieval (IR) algorithm. The Figure 2 depicts a system where a user's spoken request is first processed by a recogniser. Another possibility is to input textual requests. When phonetic indexing is used, the typed word or sentence has to be pre-processed in order to get its phonetic transcription.

After the query Q has been formed from a textual or spoken request, it is compared to each archived document D_i (i.e. to the associated metadata). This is based on a score $S(Q,D_i)$ reflecting how *relevant* is D_i with respect to Q (i.e. how likely will D_i satisfy the user's request). These relevance scores are finally used for ranking the documents, in order to output the most relevant ones.

Standard text retrieval methods, e.g. based on the vector space IR model [17] or probabilistic IR models [18], can be applied to SDR. The use of classical text retrieval methods is relevant if the indexing data consists of word transcriptions (obtained from continuous speech recognition) [13]. However, on the contrary to text retrieval, errors are introduced by the ASR systems in the document transcriptions (and in the user's input, in the case of spoken queries). Moreover, the ASR system can provide useful information in addition to the

recognized symbols, such as its error statistics or the acoustic scores resulting from the decoding process. Classical text retrieval methods do not exploit this ASR specific information. Some IR refinements are required to take these data into account, especially when phones are used as indexing terms, since phone recognition implies high error rates.

In the recent years, different works have proposed SDR methods that integrate some metadata provided by the ASR systems [7],[8],[10],[11]. As it has been mentioned in the previous section, phone confusion statistics (produced by testing the ASR system on an evaluation database) are provided by the MPEG-7 SpokenContent description. In the following, we will propose a method to take this information into account in the retrieval process.

2.2 Retrieval Model

Our IR model is based on the well-known vector space model (VSM), widely used in the traditional IR field [17]. The model creates a space in which both documents and queries are represented by vectors. Given a query Q and a document D, two T-dimensional vectors q and d are generated, where T is the pre-defined number of indexing terms. Each component of q and d represents a weight associated to a particular indexing term. Different weighting schemes can be used. The most straightforward is a binary weighting, in which a vector component is simply set to "1" if the corresponding indexing term is present. For a given term t, the corresponding components in q and d are:

$$q(t)=\begin{cases}1 & \text{if } t \in Q \\ 0 & \text{otherwise}\end{cases} \quad \text{and} \quad d(t)=\begin{cases}1 & \text{if } t \in D \\ 0 & \text{otherwise}\end{cases}. \quad (1)$$

A measure of similarity between Q and D is then estimated by using the inner product of q and d:

$$S(q,d)=\frac{1}{\|q\|.\|d\|}\sum_{t \in Q}q(t).d(t). \quad (2)$$

This similarity score reflects how relevant is the document D for a given query Q. It allows to create a list of relevant documents, ordered according to their relevance scores, which can be returned to the user.

2.3 Phone N-Grams

The indexing terms used in this study are phone N-grams [8], i.e. the sequences of N successive phones that can be extracted from the spoken content descriptions of documents and queries. In that case, the indexing terms t mentioned in Equation (1) are all the N-phone sequences extracted from the transcriptions or lattices used for indexing the queries and the documents.

For the baseline system, we have chosen $N=3$ because it has been shown to be a good compromise [8],[19].

In this paper, two approaches are proposed to improve this baseline approach. The first one is the combination of phone N-grams of different lengths. The other consists of expanding the document representation by using phone confusion probabilities. These methods are described in the two following sections. They will be compared and combined in the experiments of section 3.

2.4 Combination of N-grams Lengths

In a previous work [19], we examined the possibility of combining N-grams of different lengths. In that case, the retrieval system handles different sets of indexing terms, each one corresponding to a length N. For a given document, the retrieval scores obtained using each set separately can be combined to get a single score.

A simple combination of monogram ($N=1$), bigram ($N=2$) and trigram ($N=3$) indexing terms can be defined by using the following relevance score:

$$S_{1,2,3}(q,d)=\frac{1}{6}\sum_{N=1}^{3}N\,S_N(q,d), \quad (3)$$

where S_N represents the relevance score of Equation (2), obtained with the set of N-gram indexing terms.

This combination allows to take short indexing units into account. At the same time, it gives more weight to the longer ones, which are more sensitive to recognition errors (a single erroneous phone modifies the whole indexing term) but contain more information.

2.5 Expansion based on Confusion Probabilities

As mentioned before, the MPEG-7 SpokenContent description contains a phone confusion matrix. The elements $P(\varphi_1|\varphi_2)$ of this matrix represent the probability that the phone φ_1 is recognized instead of φ_2. The diagonal of the matrix consists of the probabilities $P(\varphi|\varphi)$ that a phone is correctly recognized.

The probability of confusion between two N-grams t and u is roughly estimated by:

$$P(u|t)=\prod_{i=1}^{N}P(\beta_i|\alpha_i). \quad (4)$$

In particular, $P(t|t)$ is an estimation of the probability that N-gram t has been correctly recognized. These probabilities can be used as weights in the calculation of the relevance score. Equation (2) is rewritten as follows:

$$S_{Conf}(q,d)=\sum_{t \in Q}P(t|t).q(t).d(t). \quad (5)$$

Reliably recognized terms thus receive higher weights. In [20], we proposed to refine Equation (5) as follows:

$$S_{Exp}(q,d)=\sum_{t \in Q}P(u_t|t).q(t).d(u_t), \quad (6)$$

where

$$u_t = \begin{cases} t & \text{if } t \in Q \cap D \\ \arg\left[\max_{t' \in D} P(t'|t)\right] & \text{if } t \in Q \cap \overline{D} \end{cases}. \qquad (7)$$

Contrary to the definition of Equation (5), the terms t that are present in Q but not in D ($t \in Q \cap \overline{D}$) are taken into account. In that case, we first determine in D which term t' has the highest probability of confusion with t. In the calculation of the relevance score, we then use $d(t')$ instead of $d(t)$ (which is null in this case) and $P(t'|t)$ instead of $P(t|t)$.

This can be seen as a query dependent expansion of the document representation, where the terms contained in $Q \cap \overline{D}$ are approximated by the terms that are the most similar to them in D.

3   EXPERIMENTS

This section reports results of SDR experiments performed on a database of German spoken documents. We evaluate the retrieval approaches described in the previous section.

3.1 Database

Experiments have been conducted with data from the PhonDat corpora (PhonDat 1 and 2) [16]. They consist of sentences read by more than 200 German speakers. We built a database of 19306 spoken documents (discarding short utterances of alphanumerical characters) that we indexed as described in section 1.

To form a set of evaluation queries, we chose 10 city names. They are listed in Table 1, along with the number of spoken documents they occur in.

Name	# Relevant Docs	Name	# Relevant Docs
Augsburg	112	Muenchen	384
Dortmund	192	Oldenburg	112
Frankfurt	368	Regensburg	448
Hamburg	528	Ulm	310
Koeln	256	Wuerzburg	96

Table 1: The 10 city names used as queries.

The phonetic transcriptions of these queries (obtained from the BOMP German pronunciation dictionary [21]) were input to our SDR system.

3.2 Evaluation Measure

Two popular measures for retrieval effectiveness are *Recall* and *Precision*. Given a set of retrieved documents, the recall rate is the fraction of relevant document in the whole database that have been retrieved:

$$\text{Recall} = \frac{\text{Number of Relevant Retrieved Doc.}}{\text{Number of Relevant Doc. in the Database}}. \qquad (8)$$

The precision rate is the fraction of retrieved documents that are relevant:

$$\text{Precision} = \frac{\text{Number of Relevant Retrieved Doc.}}{\text{Number of Retrieved Doc.}} \qquad (9)$$

The precision and recall rates depend on how many documents are kept to form the *n*-best retrieved document set. Precision and Recall vary with *n*, generally inversely with each other. To evaluate the ranked list, a common approach is to plot Precision against Recall after each retrieved document. To facilitate the evaluation of the SDR performance across different queries (each corresponding to a different set of relevant documents), we will use the plot normalization proposed by TREC [22]: the precision values are interpolated according to 11 standard Recall levels (0.0, 0.1, …, 1.0) as represented on Figure 3. These values can be averaged over all queries.

We will also evaluate the retrieval performance by means of a single performance measure, called *mean average precision* (mAP), which is the average of precision values across all recall points. It can be interpreted as the area under the Precision-Recall curve. A perfect retrieval system would result in a mean average precision of 100% (mAP = 1).
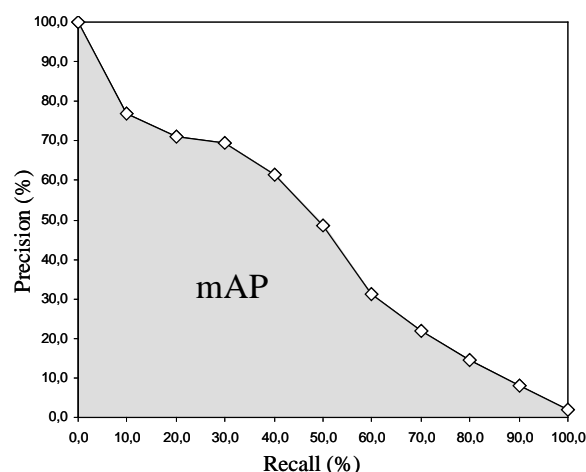

Figure 3: Precision-Recall plot, with mAP measure.

3.3 N-Gram Combination vs. Expansion

The Figure 4 represents the mAP values obtained with four different retrieval methods for each of the 10 city names used as queries. The right-most part of the figure gives these 4 measures averaged over all queries:

- The baseline performance (□) was obtained using trigrams (*N*=3) as indexing terms and relevance scores computed as described in Equation (2).
- The second mAP value (▨) results from the combination of monogram (*N*=1), bigram (*N*=2) and trigram (*N*=3) indexing terms described in section 2.4.
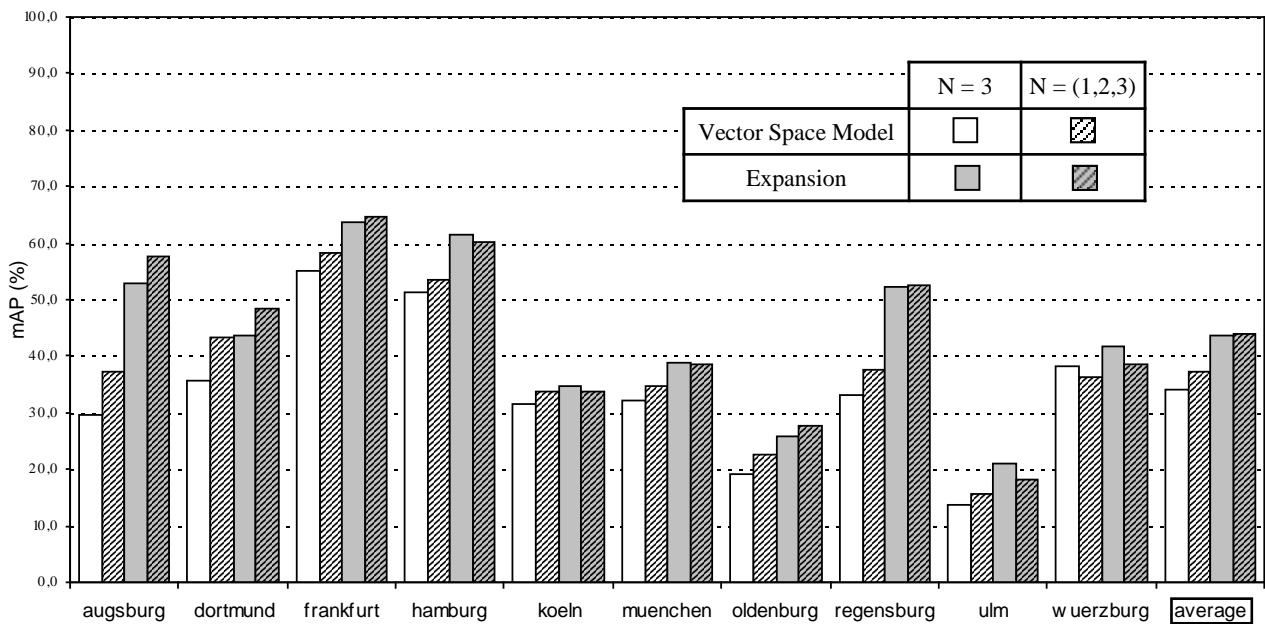
Figure 4: SDR performance measures for 10 queries and different retrieval strategies.

- The third performance measure (□) corresponds to the use of trigrams (*N*=3) with the expansion technique presented in section 2.5.
- The last case (▨) combines the 2 previous ones. It is addressed in the next section.

On average, the combination of *N*-grams increases the mean average precision from mAP=33.97% (□) to mAP=37.29% (▨) (right-most part of figure 4).

However, if we look at individual query results, there is a case (*Wuerzburg*) where the combination of 3-grams with shorter indexing terms decreases the retrieval efficiency. Taking too many indexing terms into account (the number of 1-, 2- and 3-grams extracted from a lattice can be high) can thus have a noise effect and result in a drop in retrieval performance.

The expansion method (▦) brings in much better results. It outperforms the baseline system as well as the *N*-gram combination approach for all queries.

In comparison to baseline measures (□), the mAP increase spans from 9% (*Wuerzburg*) to more than 78% (*Augsburg*). And in comparison to the combination of *N*-grams (▨), from nearly 0.5% (*Dortmund*) to more than 41% (*Augsburg*).

This expansion method compensates for certain recognition errors by taking into account some document indexing 3-gram terms that, although not contained in the query, are "closed" to them in terms of confusion probability. Even in the case of poorly performing queries (e.g. the short, three-phone-long *Ulm* query), the retrieval performance is significantly improved in comparison to both baseline (+54% for *Ulm*) and *N*-gram combination (+35% for *Ulm*) approaches.

On average, we obtained a mAP measure of 43.59% (▦) with this technique, which represents a relative increase of 17% over the *N*-gram combination approach (▨) and of 28% over baseline (□).

3.4 Combination of Both Approaches

The last mAP value (▨) results from the combination of the two previous approaches. Within each set of indexing terms, each one corresponding to a single *N* value (i.e. 3 sets in our case: *N* = 1, 2 and 3), we computed a relevance score as in Equation (6). The three resulting scores were then combined as in Equation (3).

On average (right-most part of Figure 4), this technique yields a mean average precision of 43.98% (▨). This represents only a very slight improvement (+0.9%) over the average retrieval performance obtained with the expansion technique alone (mAP=43.59%) (▦).

Applying the document expansion approach to the 3 sets of indexing terms (1-, 2- and 3-grams) simultaneously can impair the retrieval efficiency in some cases. We can observe on Figure 4, that for half of the queries, the expansion technique applied to the 3-gram set alone (▦) performs slightly better.

4 CONCLUSIONS

This paper has described a spoken document retrieval system, based on phone recognition and conform to the MPEG-7 SpokenContent standard.

Different retrieval methods have been tested, all based on the vector space model and the extraction of phone *N*-grams.

A baseline retrieval model, using phone 3-grams as indexing terms, was first defined. We then proposed two ways of refining this baseline approach.

The first one consists of combining the retrieval scores resulting from the sets of 1-, 2- and 3-grams extracted from the spoken content lattices used to index the documents. The second one is a technique to expand the vectorial representation of documents by means of phone confusion probabilities, in order to compensate for the inaccuracy of the phone recognition system.

Both methods improve the average retrieval performance in comparison to the baseline system. The expansion technique outperforms clearly the combination of *N*-grams.

Finally, we combined these two approaches. Applying the expansion method to different sets of *N*-grams (1-, 2- and 3-grams) simultaneously results in more complexity and computation cost without yielding any obvious performance improvement. In the future, we should investigate the possibility of taking into account the 1- and 2-gram indexing terms, while applying the expansion technique to 3-grams only. Other weights or combination techniques could then be considered to perform the fusion of these different information sources.

REFERENCES

[1]   MPEG Homepage: www.chiariglione.org/mpeg.

[2]   Chang S.-F., Sikora T. & Puri A., "Overview of the MPEG-7 Standard", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 688-695, June 2001.

[3]   Manjunath B.S., Salembier P., Sikora T. *et al.*, "*Introduction to MPEG-7*", Wiley, 2002.

[4]   Quackenbush S. & Lindsay A., "Overview of MPEG-7 Audio", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 725-729, June 2001.

[5]   ISO/IEC FDIS 15938-4:2001(E), *Information Technology - Multimedia Content Description Interface - Part 4: Audio*, June 2001.

[6]   Charlesworth J. P. A. & Garner P. N., "SpokenContent Representation in MPEG-7", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 730-736, June 2001.

[7]   Ferrieux A. & Peillon S., "Phoneme-Level Indexing for Fast and Vocabulary-Independent Voice/Voice Retrieval", *ESCA Tutorial and Research Workshop (ETRW), "Accessing Information in Spoken Audio"*, Cambridge, UK, April 1999.

[8]   Ng K., "*Subword-based Approaches for Spoken Document Retrieval*", PhD Thesis, Massachusetts Institute of Technology (MIT), Cambridge, MA, February 2000.

[9]   Larson M. & Eickeler S., "Using Syllable-based Indexing Features and Language Models to improve German Spoken Document Retrieval", *ISCA, Eurospeech 2003*, pp. 1217-1220, Geneva, September 2003.

[10]   Wechsler M., "*Spoken Document Retrieval Based on Phoneme Recognition*", PhD Thesis, Swiss Federal Institute of Technology (ETH), Zurich, 1998.

[11]   Srinivasan S. & Petkovic D., "Phonetic Confusion Matrix Based Spoken Document Retrieval", *23rd Annual ACM Conference on Research and Development in Information Retrieval (SIGIR'00)*, pp. 81-87, Athens, Greece, July 2000.

[12]   Rabiner L., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, February 1989.

[13]   James D. A., "*The Application of Classical Information Retrieval Techniques to Spoken Documents*", PhD Thesis, University of Cambridge, Speech, Vision and Robotic Group, Cambridge, U.K., February 1995.

[14]   SAMPA (Speech Assessment Methods Phonetic Alphabet) for German: www.phon.ucl.ac.uk/home/sampa/german.htm.

[15]   HTK (Hidden Markov Model Toolkit): http://htk.eng.cam.ac.uk/.

[16]   Bavarian Archive for Speech Signals (BAS): http://www.phonetik.uni-muenchen.de/Bas/.

[17]   Salton G. & McGill M. J., "Introduction to Modern Information Retrieval", McGraw-Hill, New York, 1983.

[18]   Crestani F., Lalmas M., Van Rijsbergen C. J. & Campbell I., ""Is This Document Relevant? ...Probably": A Survey of Probabilistic Models in Information Retrieval", *ACM Computing*

Surveys, vol. 30, no. 4, pp. 528-552, December 1998.

[19]   Moreau N., Kim H.-G., Sikora T., "Combination of Phone N-Grams for a MPEG-7-based Spoken Document Retrieval System", submitted to *EUSIPCO 2004*.

[20]   Moreau N., Kim H.-G., Sikora T., " Phonetic Confusion Based Document Expansion for Spoken Document Retrieval", submitted to *SIGIR 2004*.

[21]   Bonn Machine-Readable Pronunciation Dict.: www.ikp.uni-bonn.de/dt/forsch/phonetik/bomp.

[22]   TREC, "Common Evaluation Measures", *NIST, 10th Text Retrieval Conference (TREC 2001)*, pp. A-14, Gaithersburg, Maryland, USA, November 2001.