COMBINATION OF PHONE N-GRAMS FOR A MPEG-7-BASED SPOKEN DOCUMENT RETRIEVAL SYSTEM

Nicolas Moreau, Hyoung-Gook Kim, and Thomas Sikora

Department of Communication Systems, Technical University of Berlin
Einsteinufer 17, D-10587 Berlin, Germany (Europe)
phone: +49-30-314 28 218, fax: +49-30-314 22 514, email: moreau@nue.tu-berlin.de
web: www.nue.tu-berlin.de/wer/moreau

ABSTRACT

In this paper, we present a phone-based approach of spoken document retrieval (SDR), developed in the framework of the emerging MPEG-7 standard. The audio part of MPEG-7 aims at standardizing the indexing of audio documents. It encloses a *SpokenContent* tool that provides a description framework of the semantic content of speech signals. In the context of MPEG-7, we propose an indexing and retrieval method that uses phonetic information only and a vector space IR model. Different strategies based on the use of phone *N*-gram indexing terms are experimented.

1. INTRODUCTION

Among the multimedia documents that are today available in profusion on Internet or in private archives, many contain spoken parts. These speech signals enclose information that can be used for indexing and retrieving 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 (full transcription or manual annotation with a series of spoken keywords). A classical text retrieval system could then exploit this information. In real word applications however, hand indexing of the spoken audio material is impracticable, owing to the huge volume of most databases.

An alternative is the automatization of the transcription process by means of an automatic speech recognition (ASR) system. Due to the progress of the computation power, the ASR algorithms have now reached sufficient levels of performance that make them useable in many commercial products, from interactive vocal services to dictation programs.

The emerging MPEG-7 standard [1,2] –also called the Multimedia Content Description Interface– is an effort of the MPEG (Moving Picture Experts Group) to provide a unified and standardized way of describing the content of multimedia documents. The audio part of MPEG-7 contains a SpokenContent high-level tool [3] that provides a standardized description of the content extracted by ASR systems from spoken documents.

Section 2 will describe the MPEG-7 SpokenContent description and the indexing of spoken documents. In section 3 we present the retrieval method that will be evaluated in the experiments of section 4. Section 5 will finally give some perspectives for future investigations.

2. SPOKEN CONTENT INDEXING

Basically, the MPEG-7 SpokenContent tool defines a standardized description of the lattices (i.e. oriented graphs whose different links represent recognized terms) delivered by a recogniser. The Figure 1 gives an illustration of what an MPEG-7 SpokenContent description of the speech input "*Film on Berlin*" can be.

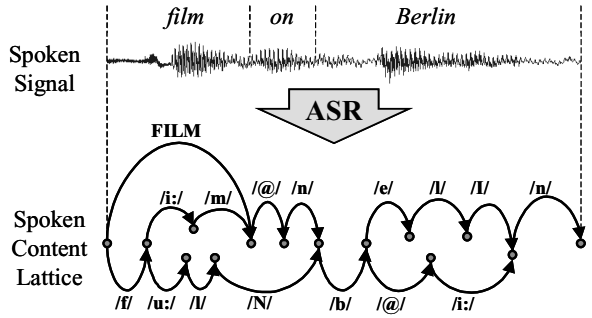

Figure 1: MPEG-7 spoken content description of an input spoken signal "*Film on Berlin*".

Each lattice link is assigned a label and the acoustic score delivered by the ASR system. The SpokenContent description also contains some additional information, such as: a word lexicon (if words are used), a phone lexicon, a phone confusion matrix and other segmental information (e.g. the speaker identity).

The standard defines two types of lattice links: word and phone. An 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.

In this study, we will only use phone graphs. Word-based indexing methods require to know *a priori* a set of keywords (keyword spotting), or to train a large vocabulary continuous speech recognition system based on a complex language model (LM). The use of sub-word units as indexing terms restrains the size of the indexing lexicon to a few dozens of units and requires no pre-defined vocabulary. However, phone recognition systems have to cope with high error rates (typically around 40%). In this case, the challenge is to exploit efficiently the MPEG-7 SpokenContent description to compensate for these high error rates.

2.1 Acoustic Models

The language used in this study is German. We used a set of 42 phone symbols derived from the 46 German phones of the SAMPA alphabet [4].
Each phone, along with the speech pause, is modeled 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 [5] was used to train the HMMs on the German "Verbmobil I" (VM I) corpus [6]. It is a large speech database consisting of spontaneous (non-prompted) speech from many different speakers and environments.

2.2 Phone Recogniser

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) trained 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 [7], 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".

3. RETRIEVAL

3.1 Retrieval Model

Our retrieval model is based on the well-known vector space model (VSM) [8]. 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 total number of possible 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:

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

The inner product of q and d is then used to estimate a measure of similarity between the query Q and the document 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 rank the documents, ordered according to their relevance scores.

3.2 Phone *N*-grams

The indexing terms used in this study are phone N-grams [9], 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 phone transcriptions or the phone lattices used to index the queries and the documents.
The question is to know what size N should be considered. If N is too small the risk is to lose the sequential information. If N is too large, the number of common indexing terms in Q and D may be too low, due to the high phone error rate. As we will see in section 4.3 the choice of N=3 seems to be optimal.

3.3 Combination of *N*-grams Lengths

In this work, we will also examine 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.
We have tried to combine monogram (N=1), bigram (N=2) and trigram (N=3) indexing terms. We obtained the final (Q,D) relevance scores through a simple linear combination of the three resulting measures of similarity:

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

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.

4. EXPERIMENTS

This section reports SDR results obtained on a database of German spoken documents.

4.1 Database

Experiments have been conducted with data from the PhonDat corpora (1&2) [6]. 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 2. The set of evaluation queries consists of 10 city names: Augsburg, Dortmund, Frankfurt, Hamburg, Koeln, Muenchen, Oldenburg, Regensburg, Ulm, Wuerzburg. A set of relevant documents corresponds to each one (between 96

and 528 documents, depending on the query). The phonetic transcriptions of these queries were input to our SDR system.

4.2 SDR Evaluation

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

$$Recall = \frac{Number\ of\ Retrieved\ Relevant\ Doc.}{Number\ of\ Relevant\ Doc.\ in\ the\ Database}. \quad (4)$$

Precision is the fraction of retrieved documents that are relevant:

$$Precision = \frac{Number\ of\ Retrieved\ Relevant\ Doc.}{Number\ of\ Retrieved\ Doc.}. \quad (5)$$

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 used the plot normalisation proposed by TREC [10]: the precision values are interpolated according to 11 standard Recall levels (0.0, 0.1, …, 1.0) as represented on Figure 2. These values can be averaged over all queries.

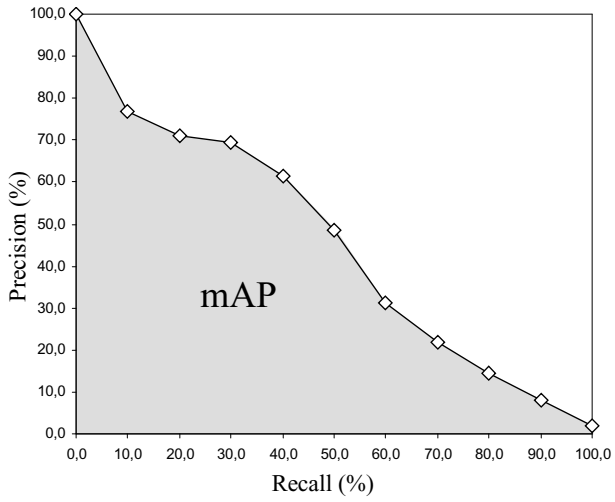

Figure 2. Precision-Recall plot, with mAP measure.

Finally, we 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).

4.3 Optimal *N*-gram Length

The Figure 3 depicts the average retrieval performance obtained with lattices and 4 different *N*-gram lengths (N= 1,2, 3 and 4).

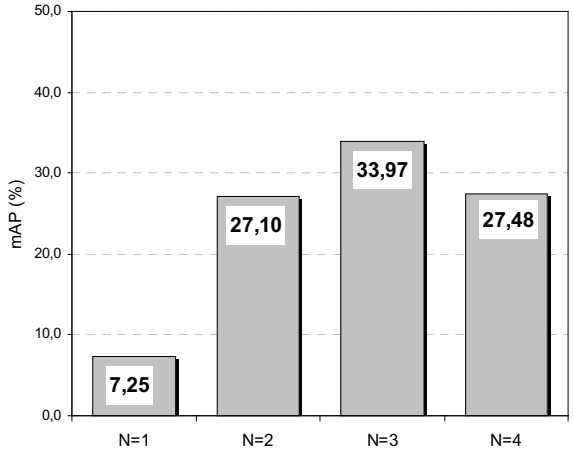

Figure 3: mAP values for different *N*-gram lengths.

The use of trigrams (*N*=3) represents the best choice. In the following section, we will examine if the combination of trigrams with bigrams (*N*=2) and individual phones (*N*=1) yields any improvement.

4.4 1-Best Transcriptions vs. Lattices

Figure 4 represents the mAP values obtained for each query with different indexing strategies. The right-most part gives the mAP values averaged over all queries.

The 2 first measures were obtained with *N*=3. For the first one (□), only the 1-best transcriptions delivered by the ASR system were used for indexing the documents. The second (■) is obtained with lattices. In any case, the use of lattices brings in an improvement. Lattices represent an expansion of the 1-best transcriptions. It takes into account alternative phone hypotheses which allow to recover correctly recognised 3-grams which were not part of the best sequence.

As expected, lattices yield better results than 1-best transcriptions for all queries. On average, the mean average precision increases from mAP= 28.89% with transcriptions (□) up to mAP= 33.97% with lattices (■).

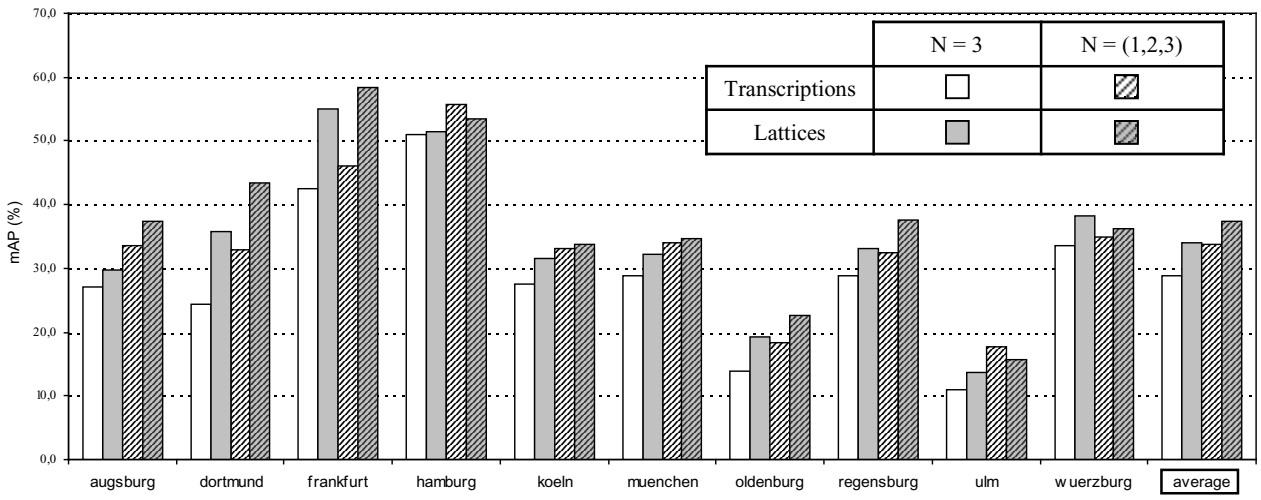

Figure 4: Compared SDR performance measures with different indexing and retrieval strategies.

4.5 Combination of *N*-gram lengths

The two other measures displayed on Figure 4 are obtained with 1-best transcriptions (▨) and lattices (▨), using the combination of 1-, 2- and 3-grams described in equation (3).
With 1-best transcriptions, the combined multigram approach yields an improvement in any case, compared to the use of 3-grams. On average, we obtained mAP=28.89% with 3-grams (□) and mAP=33.86% with the combined approach (▨).
With lattices, the combination of 3-grams with shorter indexing terms decreases the retrieval efficiency for one query ("Wuerzburg"), compared to the use of 3-grams. 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.
But it should be noticed that, on average, the combined approach also results in a global performance improvement in the case of lattices. We thus obtained an average mAP of 37.29% (▨), which is better than the 33.97% average mAP obtained with 3-grams (■).

5. CONCLUSION

This paper presented a German spoken document indexing and retrieval system, based on phone lattices and conform to the MPEG-7 SpokenContent standard.
Several indexing and retrieval approaches were compared.
With a simple baseline retrieval model based on phone 3-grams, we could verify that the indexing of spoken documents with lattices outperforms the use of simple transcriptions. We then proposed a retrieval approach combining 1-, 2- and 3-grams that improves the average retrieval performance in comparison to the baseline system, whichever indexing method is used (1-best transcriptions or lattices).
These experiments constitute a first milestone in the development of our phone-based German SDR system.
Several other data enclosed into the MPEG-7 SpokenContent descriptions can be used to improve further the retrieval efficiency. In a future study, we will use phone confusion probabilities to expand the representations of documents in the vector space model, thus trying to compensate for the inaccuracy of the phone recognition system.

REFERENCES

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

[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] 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.

[4] SAMPA (Speech Assessment Methods Phonetic Alphabet): www.phon.ucl.ac.uk/home/sampa.

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

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

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

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

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

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