Phonetic Confusion Based Document Expansion for Spoken Document Retrieval

Nicolas Moreau, Hyoung-Gook Kim, Thomas Sikora

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

Abstract

This paper presents a phone-based approach of spoken document retrieval (SDR), developed in the framework of the emerging MPEG-7 standard. We describe an indexing and retrieval system that uses phonetic information only. The retrieval method is based on the vector space IR model, using phone *N*-grams as indexing terms. We propose a technique to expand the representation of documents by means of phone confusion probabilities in order to improve the retrieval performance. This method is tested on a collection of short German spoken documents, using 10 city names as queries.

1. Introduction

Huge amounts of audio-visual documents are today available on Internet or in private archives. The audio streams of these documents often contain spoken parts that enclose a lot of semantic information. The extraction of this information has become a key challenge for the development of efficient methods to index and retrieve audio-visual documents.

One method of exploiting the spoken information is to have a human listen and transcribe it into textual information. However, hand indexing is impracticable, owing to the huge volume of most spoken audio 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 a continuous speech context.

This study presents an ASR-based system for the indexing and retrieval of German spoken documents. It was developed in accordance with the MPEG-7 standard [1], which aims at providing a unified way of describing the content of multimedia documents.

Our indexing system is vocabulary independent. It extracts phonetic information from speech through an ASR system. This process implies that a lot of recognition errors are introduced within the indexing description. In this context, specific retrieval methods are required.

In the past, several works have proposed spoken document retrieval (SDR) approaches that use some ASR specific information [2][3], in particular the phone confusion matrix, to compensate for the imprecision of the indexing process. In the same way, the SDR system presented here exploits the phone confusion statistics in order to expand the phonetic representation of the documents.

The paper is structured as follows. Section 2 describes the indexing system. The benefits and drawbacks of indexing spoken documents with phones are discussed. In section 3, we present our baseline retrieval model and propose to improve it by taking phone confusion statistics into account. Several experimental results are reported and commented in section 4.

2. Spoken content indexing

2.1. MPEG-7 SpokenContent description

The audio part of MPEG-7 contains a *SpokenContent* high-level tool [4] that provides a standardized description of the information extracted by ASR systems. Basically, this description consists of a lattice of recognition hypotheses (i.e. an oriented graph whose different links represent recognized terms). Each lattice link is assigned a label and the acoustic score delivered by the ASR system. The standard defines two types of lattice links: word and phone. A MPEG-7 lattice can thus be a word-only graph, a phone-only graph (as in our case) or combine word and phone hypotheses.

The SpokenContent description also contains some additional information: the word or phone lexicon, a series of phone confusion statistics and other segmental information (e.g. the speaker identity).

2.2. Phone based indexing

Many SDR systems proposed in the past consisted in linking a word-level ASR engine with a traditional text retrieval system [5]. In the case of word-based indexing, the vocabulary has to be known beforehand, which precludes the handling of out-of-vocabulary (OOV) words. Furthermore, the derivation of complex language models is necessary for reasonable quality LVCSR systems (large vocabulary continuous speech recognition). This requires huge amounts of training data that contain several occurrences of recognition-vocabulary words.

In the recent years, other approaches have considered the indexation of spoken documents with sub-lexical units instead of word hypotheses [2][3][6]. In that case, a limited amount of sub-word models is necessary and any sentence can be indexed with a sequence –or a lattice– of sub-lexical indexing terms, such as phones, phonemes or syllables.

In this study, we will only use phone graphs. The indexing system we developed does not require any *a priori* set of keywords (as in keyword spotting systems), nor complex language models (as in LVCSR systems). 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 to cope with high error rates (typically around 40%). In our case, the challenge is to exploit efficiently the MPEG-7 SpokenContent description to compensate for these high error rates.

Besides, the use of only phonetic indexing might lose discrimination power between relevant and irrelevant

documents when compared to word indexing, because of the exclusion of lexical knowledge. However this study focuses on a simple SDR task with short documents and single word queries (one can imagine the scenario of a database of photos annotated with short spoken descriptions). In that case, we think that the use of a simple, vocabulary-independent phone recogniser is a reasonable indexing approach.

2.3. Phone recognizer

The language considered in this study is German. We used a set of 42 phones modeled by context independent HMMs 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. The 39-dimensional observation vectors consist of 12 mel-frequency cepstral coefficients (MFCCs), the energy, and the first and second derivatives. The HTK toolkit was used for training the HMMs on the German "Verbmobil I" (VM I) corpus, a large collection of spontaneous speech from many different speakers.

Phone recognition is performed without any lexical constraints. The phone HMMs are looped according to a bigram language model (LM), which was trained from the transcriptions of the "Vermobil II" corpus. Given a spoken input, our ASR system produces an output phone lattice containing several hypothesized phonetic transcriptions. The set of indexing symbols was reduced by merging some acoustically similar phone classes. We mapped our 42 phones to 32 German "phonemes" as proposed by [3]. For more convenience, we will continue 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. We obtained a phone error rate (PER) of 43.0%. This test also provided the phone confusion probabilities which are enclosed, along with the phone lexicon and the phone lattices, in the MPEG-7 SpokenContent descriptions.

3. Retrieval

3.1. Retrieval model

Our IR model is based on the well-known vector space model (VSM). 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 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:

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

There exists more refined weighting methods that take into account term frequencies in *Q* and *D*, and/or the inverse document frequency. However, the weighting issue will not be addressed here.

Finally, the inner product of *q* and *d* is used for estimating a measure of similarity between the query *Q* and the document *D*:

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

This similarity score reflects how relevant *D* is, with respect to *Q*. A list of relevant documents, ordered according to their scores, is then returned to the user.

3.2. Indexing with phone *N*-grams

As in [2], the indexing terms used in this study are phone *N*-grams, i.e. the sequences of *N* successive phones present in the spoken content descriptions of documents and queries. In that case, the set of indexing terms *t* used in equation (1) consists of all the *N*-phone sequences found along the different paths of the indexing lattices. In the following experiments, we will also consider the case where *N*-grams are extracted from the best path only (i.e. from the best document transcription). According to the results of a previous study [7], we set *N*=3.

3.3. Weighting with 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 phone φ_1 is recognized instead of φ_2. The diagonal of the matrix consists of the probabilities $P(\varphi|\varphi)$ that a phone φ is correctly recognized. The probability of confusion between two *N*-grams *t* and *u* is roughly estimated by:

$$P(u|t)=\prod_{i=1}^{N} P(\beta_i\,|\alpha_i)\,. \qquad (3)$$

where α_i and β_i are the i^{th} phones of *t* and *u* respectively. The values $P(t|t)$ can be considered as the probability that *N*-gram *t* has been correctly recognized. The $P(t|t)$ probabilities can be used as weights in the calculation of the relevance score. The equation (2) is rewritten as follows:

$$S_{Conf}(q,d)=\sum_{t\in Q} P(t|t).q(t).d(t)\,. \qquad (4)$$

The terms that can be considered as the most reliably recognized by the ASR system thus receive the highest weights. It should be noted that the normalizing factor of equation (2) has been discarded in this case. We empirically verified that this method performs better without normalising the scores.

3.4. Expansion of the Representation

The basic idea is to redefine the calculation of the relevance score as proposed in equation (4) by taking new elements of the document vector *d* into account. The new definition of the relevance score is the following:

$$S_{Exp}(q,d)=\sum_{t\in Q} P(u_t|t).q(t).d(u_t)\,, \qquad (5)$$

where

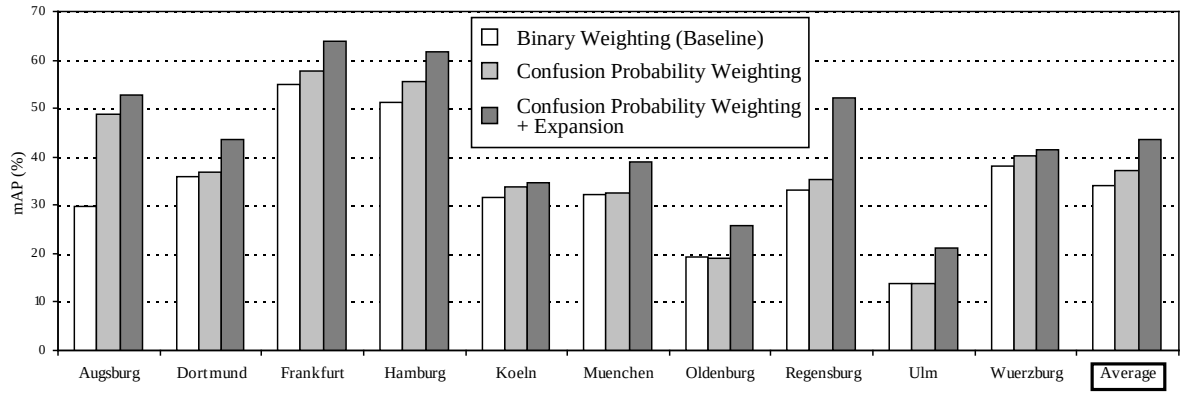

Figure 1: SDR performance measures for 10 different queries and 3 weighting strategies.

$$u_t = \begin{cases} t & \text{if } t \in Q \cap D \\ \arg\left[\max_{t' \in D} P(t'|t) \right] & \text{if } t \in Q \cap \overline{D} \end{cases} . \quad (6)$$

Contrary to the definition of equation (4), the terms t that are present in Q but not in D (i.e. $t \in Q \cap \overline{D}$) 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 $d(t')=1$ instead of $d(t)$ (null in this case) and $P(t'|t)$ instead of $P(t|t)$.

A query expansion technique using phone confusion is proposed in [2]: given a query N-gram term t, a fixed-length list of terms having a high probability of confusion with t is added to Q, independently from the document collection.

The method proposed here expands the representation of each document individually, according to a given query. The terms contained in $Q \cap \overline{D}$ are approximated by the terms that are the most similar to them in D.

4. Experiments

4.1. Database

Experiments have been conducted with data from the PhonDat corpora (PhonDat 1 and 2) consisting of short 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 average length of the documents is 3.9 seconds (37.7 phones per transcription on average).

The set of evaluation queries consists of 10 city names: Augsburg, Dortmund, Frankfurt, Hamburg, Koeln, Muenchen, Oldenburg, Regensburg, Ulm, Wuerzburg. In a word-based indexing approach, one can imagine that these proper names would be out-of-vocabulary words. A set of relevant documents corresponds to each query (between 96 and 528 documents, depending on the query). Their phonetic transcriptions were used as single word queries.

4.2. Evaluation

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

The precision and recall rates depend on how many documents are kept to form the n-best retrieved document list. They 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. We used the plot normalization proposed by TREC [8].

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. Document Expansion

Figure 1 represents the mAP values obtained with each of the 10 city names used as queries and different weighting strategies. The right-most part of the figure represents the average values of the 10 query specific results.

The first measure (□) was yielded from the baseline system, i.e. using the binary weighting and a simple inner vector product (equation 2) to compute the IR similarity measure. The second one (▒) results from the use of confusion probabilities to weight the terms of the inner vector product (equation 4). The last mAP measure (■) is obtained with the expansion technique described in equation (5).

The exact mAP values for each of the 10 queries are reported in Table 1. The last line gives the averaged values. In the second and third columns, the relative performance improvements in comparison to the baseline system (first column) are given enclosed in parentheses. These results are commented in the following sections.

4.3.1. Weighting with Confusion Probabilities

The first comment concerns the introduction of confusion probability weights (equation 4) into the calculation of the relevance score.

We observe on Figure 1 that not every queries benefit from the introduction of confusion probabilities (▒ compared to □). For some of them, this weighting factor brings no improvement in comparison to the performance of the baseline system (*Muenchen*, *Oldenburg*, *Ulm*).

In the other cases, the retrieval effectiveness is clearly increased. A query such as *Augsburg* might contain one or a

few 3-grams for which the ASR system have produced high recognition probabilities.

On average (right part of Figure 1), this technique improves the overall retrieval performance. In comparison to the baseline average performance (mAP=33.97%), the mAP increases by 9.8% (mAP=37.29%) as can be seen in Table 1.

Table 1: mAP values for the 10 queries.

	mAP(%)				
	Binary	Confusion		Confusion + Exp.	
Augsburg	29.66	48.73	(+64.3%)	52.91	(+78.3%)
Dortmund	35.82	36.86	(+2.9%)	43.50	(+21.4%)
Frankfurt	55.05	57.58	(+4.6%)	63.77	(+15.8%)
Hamburg	51.34	55.42	(+7.9%)	61.60	(+19.9%)
Koeln	31.52	33.71	(+6.9%)	34.72	(+10.1%)
Muenchen	32.13	32.40	(+0.8%)	38.93	(+21.1%)
Oldenburg	19.23	18.95	(-1.4%)	25.72	(+33.7%)
Regensburg	33.06	35.22	(+6.5%)	52.13	(+57.6%)
Ulm	13.68	13.68	(+0.0%)	21.08	(+54.0%)
Wuerzburg	38.19	40.30	(+5.5%)	41.58	(+8.8%)
Average	**33.97**	**37.29**	(+9.8%)	**43.59**	(+28.3%)

4.3.2. Expansion

The document representation expansion described in part 3.4 further improves the retrieval effectiveness.

As reported in Figure 1 (■) and Table 1, the average performance is improved by 28.3% (from mAP=33.97% to 43.59%) in comparison to the baseline system (□), and by 16.9% in comparison to the use of confusion weighting with no expansion (■).

Even in the case of poorly performing queries (in particular the *Oldenburg* and *Ulm* queries), this approach improves significantly the retrieval performance. In the case of the short, three phone long *Ulm* query, for example, we get one of the best relative mAP improvement in comparison to the baseline value (+54.0%).

Queries that are poorly retrieved indicate that the recognition of the corresponding targets within the indexed documents contain a lot of recognition errors. This is particularly problematic when the query is short. In that case, the corresponding targets may not contain any correctly recognized 3-grams that can compensate for the badly recognized ones. The use of the expansion strategy can recover some of the missed targets by taking into account some document indexing 3-grams that, although different from the ones contained in the query, are “closed” to them in terms of confusion probability.

4.4. Lattices vs. transcriptions

Compared to simple phonetic transcriptions, the use of multi-hypothesis phone lattices also represents an expansion of document representations. To verify if the combination of both expansion techniques is not redundant, we tested our method using the “1-best” transcription (i.e. the best path in the lattice) as document index. The results are reported in Table 2, along with the ones obtained with lattices (mAP values average over all queries).

As expected in the baseline case (Binary), the use of lattices yields an improvement compared to the use of simple transcriptions (+17.6%). More interesting is to observe the same phenomenon when applying the confusion-based expansion technique (Confusion + Exp): lattices perform significantly better than transcriptions in that case too (+14.7%). The application of our expansion technique to multi-hypothesis lattices seems to be relevant.

Table 2: Average mAP (%) with transcriptions and lattices.

	1-Best	Lattice
Binary	28.89	33.97
Confusion + Exp.	38.01	43.59

5. Conclusions

This paper presented a German spoken document indexing and retrieval system, based on phone lattices and in conformance with the MPEG-7 standard. The retrieval method uses a simple vector space model with phone *N*-grams as indexing units. In order to compensate for the inaccuracy of the phone recognition system, we proposed a technique to expand the representation of documents by means of phone confusion probabilities. This method was tested using 10 city names as queries. Compared to the baseline system, it improved the retrieval performance by about 28% on average. The application of this expansion technique to multi-hypothesis lattices instead of simple transcriptions proved to be relevant.

The effects of the *N*-gram length should be further studied in the future. In previous works, we already obtained some interesting results by combining *N*-grams of different lengths [7]. As query lengths have an impact on the retrieval performance, another perspective would be to consider different *N*-gram lengths for different query lengths.

Besides, this approach still has to be tested on larger databases, with more queries, and to be compared to query expansion techniques.

6. References

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

[2] Ng K. & Zue V. W., "Subword-based Approaches for Spoken Document Retrieval", *Speech Communication*, vol. 32, no. 3, pp. 157-186, October 2000.

[3] Wechsler M., Munteanu E. & Schäuble P., "New Techniques for Open-Vocabulary Spoken Document Retrieval", *SIGIR'98*, pp. 20-27, August 1998.

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

[5] James D. A., "*The Application of Classical Information Retrieval Techniques to Spoken Documents*", PhD Thesis, University of Cambridge, February 1995.

[6] Larson M. & Eickeler S., "Using Syllable-based Indexing Features and Language Models to improve German Spoken Document Retrieval", *Eurospeech'03*, pp. 1217-1220, September 2003.

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

[8] TREC, "Common Evaluation Measures", *10th Text Retrieval Conference*, pp. A-14, November 2001.