

# INTEGRATION OF METADATA IN SPOKEN DOCUMENT SEARCH USING POSITION SPECIFIC POSTERIOR LATTICES

Jorge Silva<sup>1,1†</sup>, Ciprian Chelba<sup>2,1†</sup> and Alex Acero<sup>3</sup>

<sup>1</sup>University of Southern California, jorgesil@usc.edu

<sup>2</sup>Google, Kirkland, WA, USA, ciprianchelba@google.com

<sup>3</sup>Microsoft Corporation, Redmond, WA, USA, alexac@microsoft.com

## ABSTRACT

<sup>1</sup> This paper addresses the problem of integrating speech and text content sources for the document search problem, as well as its usefulness from an ad-hoc retrieval — keyword search — point of view. Position Specific Posterior Lattices (PSPL) is naturally extended to deal with both speech and text content, where a new relevance ranking framework is proposed for integrating the different sources of information available.

Experimental results on the MIT iCampus corpus show a relative improvement of 302% in Mean Average Precision (MAP) when using speech content *and* metadata as opposed to just metadata (which constitutes about 1% of the amount of words in the transcription of the speech content).

## 1. INTRODUCTION

Ever increasing computing power and connectivity bandwidth together with falling storage costs result in an overwhelming amount of data of various types being produced, exchanged, and stored. In the context of spoken documents (SDs), the availability and usefulness of large collections is limited strictly by the lack of adequate technology to exploit them [1, 2]. Manually transcribing speech is expensive and consequently, automatic speech recognition (ASR) turns out to be the natural direction to searching and navigating SD collections.

In this direction, PSPL was proposed as a way to extend the key-word search paradigm from text documents to SDs for realistic WER scenarios [1, 2]. The approach calculates posterior probabilities of words at a given integer position — soft indexing — to model the uncertainty of the SD content and significantly reduce the size of the ASR lattice. The position information is used for incorporating proximity in the scoring paradigm by allowing the calculation of distance- $k$  skip  $n$ -gram expected counts strictly based on the inverted index.

SD collections usually have metadata or text information appended to it. On one hand, the text metadata is deterministic, very limited in size, and it very likely differs from the actual spoken transcription, which may limit its relevance to the content of the document. On the other hand, the ASR output is a noisy representation of the underlying lexical content and

therefore we need to deal with content document uncertainty. Consequently, an approach that optimally integrates these two sources of information by considering their intrinsic nature is desirable. In this work we present a framework to address this problem. First, we propose a simple method for integrating metadata and speech content for the retrieval problem. Second, we investigate how much performance gain is provided by the SD material with respect to a baseline system that uses only the text-metadata for document search.

Regarding the first point, this work presents a novel approach for integrating metadata and SD information in a unified framework based on the PSPL [1, 2]. The approach takes advantage of the generality of the PSPL approach to incorporate deterministic and stochastic types of document content. Based on that, a framework for integrating content type-specific scores is proposed, taking into consideration the different nature of those sources. Regarding the second point, this work presents experimental evidence supporting the fact that the SD source provides significant improvement in Mean Average Precision (MAP) with respect to the scenario where only the metadata is considered for the problem. Surprisingly, this result is obtained using an ASR system with high WER.

## 2. POSITION SPECIFIC POSTERIOR LATTICES

Of essence to fast retrieval on static text document collections of medium to large size is the use of an *inverted index*. The inverted index stores a list of hits for each word in a given vocabulary. The hits are grouped by document. For each document, the list of hits for a given query term must include position — needed to evaluate counts of proximity types — as well as all the context information needed to calculate the relevance score of a given document [1]. If we want to extend this direction to SDs, we are faced with a dilemma. On one hand, using 1-best ASR output to be indexed is suboptimal due to the high WER, likely to lead to low precision-recall metrics [2]. On the other hand, ASR lattices do have much better WER — in our case the 1-best WER was 55% whereas the lattice WER was 30% — but the position information needed for recording a given word hit is not readily available in ASR lattices.

<sup>1†</sup>Work conducted while with Microsoft Corporation, WA, USA.

Let's consider that a traditional text-document hit for given word consists of just (document id, position). In this context, the ASR lattices do contain the information needed to evaluate proximity information, since on a given path through the lattice we can easily assign a position index to each link/word in the normal way. Each path occurs with a given posterior probability, easily computable from the lattice, so in principle one could index *soft-hits* which specify (document id, position, posterior probability) for each word in the lattice. A dynamic programming algorithm was proposed for performing this computation, details in [2, 1]. For computing forward pass one needs to split the forward probability arriving at a given node  $n$ ,  $\alpha_n$ , according to the length  $l$  of the partial paths that start at the start node of the lattice and end at node  $n$ , [1]:

$$\alpha_n[l] \doteq \sum_{\pi: \text{end}(\pi)=n, \text{length}(\pi)=l} P(\pi) \quad (1)$$

The backward probability  $\beta_n$  has the standard definition and dynamic recursion, where the dynamic recursion for  $\alpha_n[l]$  is formally presented in [1, 2]. Using these forward-backward variables the posterior probability of a given word  $w$  occurring at a given position  $l$  in the lattice can be easily calculated using:

$$P(w, l|LAT) = \sum_n \text{s.t. } \alpha_n[l] \cdot \beta_n > 0 \frac{\alpha_n[l] \cdot \beta_n}{\beta_{start}} \cdot \delta(w, \text{word}(n)) \quad (2)$$

Finally, the PSPL is a representation of the  $P(w, l|LAT)$  distribution: for each position bin  $l$  store the words  $w$  along with their posterior probability  $P(w, l|LAT)$ . In our case the speech content of a typical SD was approximately 1 hr long; speech files were segmented into shorter segments. A SD thus consists of an ordered list of segments. For each segment we generate a corresponding PSPL lattice.

### 3. SD INDEXING AND SEARCH USING PSPL

Consider a given query  $\mathcal{Q} = q_1 \dots q_i \dots q_Q$  and a SD  $D$  represented as a PSPL. The possible word sequences in the document  $D$  clearly belong to the ASR vocabulary  $\mathcal{V}$  whereas the words in the query may be out-of-vocabulary (OOV). We assume that the words in the query are all contained in  $\mathcal{V}$ ; OOV words are mapped to UNK and cannot be matched in any document  $D$ . For all query terms, a 1-gram score is calculated by summing the PSPL posterior probability across all segments  $s$  and positions  $k$ . The results are aggregated in a common value  $S_{1-gram}(D, \mathcal{Q})$ :

$$S(D, q_i) = \log \left[ 1 + \sum_s \sum_k P(w_k(s) = q_i|D) \right] \quad (3)$$

$$S_{1-gram}(D, \mathcal{Q}) = \sum_{i=1}^Q S(D, q_i) \quad (4)$$

where similar to [3], the logarithmic tapering off is used for discounting the effect of large counts in a given document.

The PSPL ranking scheme takes into account proximity in the form of matching  $N$ -grams present in the query. We calculate an expected tapered-count for each  $N$ -gram  $q_i \dots q_{i+N-1}$  in the query and then aggregate the results in a common value  $S_{N-gram}(D, \mathcal{Q})$  for each order  $N$ , Eq.(6), where the different proximity types are combined by taking the inner product with a vector of weights  $\{\lambda_N : N = 1, \dots, Q\}$ , Eq.(7).

$$S(D, q_i \dots q_{i+N-1}) = \log \left[ 1 + \sum_s \sum_k \prod_{l=0}^{N-1} P(w_{k+l}(s) = q_{i+l}|D) \right] \quad (5)$$

$$S_{N-gram}(D, \mathcal{Q}) = \sum_{i=1}^{Q-N+1} S(D, q_i \dots q_{i+N-1}) \quad (6)$$

$$S(D, \mathcal{Q}) = \sum_{N=1}^Q \lambda_N \cdot S_{N-gram}(D, \mathcal{Q}) \quad (7)$$

Only documents containing all the terms in the query are returned. In the current implementation the weights increase linearly with the  $N$ -gram order.

## 4. INTEGRATION OF METADATA

SDs rarely contain only speech. Often they have a title, author and creation date. The idea of saving context information when indexing HTML documents and web pages can thus be readily used for indexing spoken documents. PSPL lattices can be used to represent text content and consequently to naturally integrate the metadata in the search framework [1]. In this scenario there is no document content uncertainty and consequently the equivalent PSPL lattice has only one entry for every position bin with position specific probability equal to 1.0.

For representing the fact that documents may have text data in addition to the spoken information, we represent documents as collections of segments, as proposed in the previous section. However, we introduce a new attribute on those segments that allows having different segment categories. For doing that, we use different segment type labels for representing the speech content and the metadata content for a given document. Note that this categorization allows considering the different nature of those sources of information for computing the relevance ranking score. The next sub-section presents that formalization.

### 4.1. Relevance Ranking Considering Segment Types

Again let's consider a given query  $\mathcal{Q} = q_1 \dots q_i \dots q_Q$  and a SD  $D$ . To be more specific, the document  $D$  is a collection of segments denoted by  $\Theta_D$ , where  $\Theta_D$  is partitioned in different segment types, Eq.(8).

$$D \doteq \Theta_D = \cup_{k=1}^{N_D} \Theta_D^{type.k} \quad (8)$$

Based on this partition, we calculate individual scores for the different segment types,  $\forall k \in \{1, \dots, N_D\}$ , by:

$$S^{type.k}(D, \mathcal{Q}) = \sum_{N=1}^Q \lambda_N \cdot S_{N-gram}^{type.k}(D, \mathcal{Q}) \quad (9)$$

where the N-gram scores are the generalization of Eqs.(3-6) for the case of having specific segment types:

$$S^{type.k}(D, q_i \dots q_{i+N-1}) = \log [1 + \sum_{s \in \Theta_D^{type.k}} \sum_k \prod_{l=0}^{N-1} P(w_{k+l}(s) = q_{i+l}|D)] \quad (10)$$

$$S_{N-gram}^{type.k}(D, \mathcal{Q}) = \sum_{i=1}^{Q-N+1} S^{type.k}(D, q_i \dots q_{i+N-1}) \quad (11)$$

Finally the global score for the SD  $D$  is a linear combination of the segment-type specific ones using segment weights  $\{\lambda_{type.k} : k = 1, \dots, N_D\}$ , as shown in Eq. (12). This can be justified in a Bayesian framework under the natural assumption that those information sources can be considered independent. The weights in this expression provide the flexibility to adjust the global score to the nature of the segment types presented in the problem.

$$\hat{S}(D, \mathcal{Q}) = \sum_{k=1}^{N_D} \lambda_{type.k} \cdot S^{type.k}(D, \mathcal{Q}) \quad (12)$$

## 5. EXPERIMENTS

All our experiments were conducted on the iCampus corpus [4] prepared by MIT CSAIL. It consists of about 169 hours of lecture material recorded in the classroom. The corpus contains 90 Lectures (78.5 hours) and 79 Assorted MIT World seminars (89.9 hours). Each lecture comes with a word-level manual transcription that segments the text into semantic units that could be thought of as sentences. The speech style is in between planned and spontaneous recorded at a sampling rate of 16kHz (wide-band). Regarding the metadata, the corpus provides titles, abstracts and bibliography of the speakers for the Assorted MIT World seminars documents (89.9 hours). *The relative size of the metadata with respect to the spoken content, in number of transcribed words, is less than 1%.*

The 3-gram language model used for decoding is trained on a large amount of text data, primarily newswire text. The vocabulary of the ASR system consists of 110k words, selected based on frequency in the training data. The acoustic model is trained on a variety of wide-band speech and it is a standard clustered tri-phone, 3-states-per-phone model. *Neither model has been tuned in any way to the iCampus scenario.* On the first lecture L01 of the Introduction to Computer Programming Lectures the WER of the ASR system was 44.7%; the OOV rate was 3.3%. We generated 3-gram lattices and PSPL lattices using the above ASR system.

For the queries we have asked a few colleagues to issue queries using the index built from the manual transcription. We collected 116 queries. The query out-of-vocabulary rate (Q-OOV) was 5.2% and the average query length was 1.97 words. Since our approach so far does not index sub-word units, we cannot deal with OOV query words. We have thus removed the queries which contained OOV words — resulting in a set of 96 queries. For evaluation we have taken the output of a standard retrieval engine working on the segment transcriptions according to one of the TF-IDF flavors.

### 5.1. Metadata Integration Analysis

In this analysis we consider two categories of segment types for every document: segments of type `speech`, PSPL lattices generated from the ASR word lattices as presented in Section 2; and segments of type `metadata`, PSPL lattices generated directly from the text information in which we incorporate all the metadata available for the documents. In this initial experimental setting, we just consider the section of the corpus that has text-metadata available, namely the MIT World seminars documents (89.9 hours). *Our choice is thus biased in favor of the metadata-only scenario — many documents do not contain any metadata.* The purpose of this set of experiments is to analyze performance changes as a function of the `speech` - `metadata` relative weight in the scoring framework, Eq.(12). We explore different weight combinations under the following condition:  $\lambda_{type.speech} + \lambda_{type.metadata} = 1.0$ . Note that this allows one to evaluate the limit cases of using only the metadata,  $\lambda_{type.metadata} = 1.0$ , or only the speech content,  $\lambda_{type.metadata} = 0.0$ .

Table 1 presents MAP and R-precision evolution for different weight combinations. As expected, in the process of increasing the relative weight of the metadata there is an improvement in performance. Performance increases monotonically with the magnitude of the metadata weight from 1.62% to 2.4%. This trend can be explained because the metadata content is much more reliable than the speech information and highly related to the content of the associated SD. Consequently, giving higher ranking to relevant documents obtained from the metadata than from the speech side improves the ranking performance. Supporting this point, Precision for the metadata and speech content is 1.0 and 0.32, respectively.

However, when placing all the weight on the metadata segments there is a significant drop in performance. Looking at it the other way, the performance gain obtained by adding the speech content with respect to only considering the metadata is 302% relative. Consequently, adding SD information provides a dramatic gain in performance, which can be explained by the fact that the metadata constitutes only about 1% of the amount of words in the transcription speech content. The principal reason is that although the PSPL speech segments extracted from the spoken content have intrinsic uncertainty, this information is more representative of the underlying information in the SD — in this case, its transcription — than the very limited amount of text-metadata. This fact can be clearly observed in the significant difference in Recall between metadata and speech content, which is 0.056 and 0.815, respectively.

### 5.2. Performance Analysis for different Metadata Quality Conditions

We explore the retrieval performance gain brought by adding spoken content as a function of the metadata quality. In or-

| Metadata Weight     | MAP           | R-precision |
|---------------------|---------------|-------------|
| 0.0 (speech only)   | 0.6449        | 0.5905      |
| 0.1                 | 0.6554        | 0.5999      |
| 0.3                 | 0.6583        | 0.6022      |
| 0.5                 | 0.6599        | 0.604       |
| 0.7                 | <u>0.6606</u> | 0.6048      |
| 1.0 (metadata only) | <u>0.1642</u> | 0.1408      |

**Table 1.** Retrieval performance as a function of the weight placed on metadata and speech content

der to generate metadata of different relevance degrees, for a given document  $D$ , we enrich its original metadata by adding metadata segments which correspond to the transcription of some its speech segments at different sampling rates. We thus generated 4 different metadata sets: metadata + 1% transcription, metadata + 4% transcription, metadata + 8% transcription, and metadata + 10% transcription.

The entire iCampus corpus can be used for evaluation since we now have metadata for every spoken document in the corpus. For evaluating the performance using both metadata and speech, we consider the same relative weight across all these experiments:  $\lambda_{type.metadata} = 0.8$ , based on results obtained in the previous subsection.

Table 2 presents this relative improvement in MAP for the different metadata conditions. It can be seen that in all metadata scenarios there is a significant gain in performance by adding the spoken content. In particular, even when the metadata segments account for more than 10% of the transcriptions, the spoken content still provides a relative improvement of more than 200% in MAP. Surprisingly, these results are obtained using an ASR system with high WER (44.7%). We can conclude that despite limitations of current state-of-the-art ASR systems, the spoken content is an important information source to consider for the spoken document retrieval problem, even in scenarios with relatively high availability of text-metadata.

### 5.3. Comparison with Related Work

The problem of metadata and SD integration was recently addressed in [5]. One of the experimental results shows that indexing the spoken content in addition to the text metadata yielded only marginal gains in MAP.

We can mention two main reasons that would account for the differences with our conclusions regarding the relative usefulness of speech versus text meta-data for search in spoken documents: [5] used human judgments as the reference scenario whereas our work uses the output of an IR engine on the transcription of the spoken documents. Secondly, the metadata in [5] was extremely informative with regard to the document contents (prepared by trained human indexers); in many practical scenarios, including the one we addressed, the meta-data is scarce, resulting in very poor MAP performance.

| Sampl. Prob. | Meta. (MAP) | Meta.-Speech. (MAP) | Relative gain (%) |
|--------------|-------------|---------------------|-------------------|
| 0.01         | 0.106       | 0.647               | 510.1             |
| 0.04         | 0.131       | 0.647               | 394.6             |
| 0.08         | 0.182       | 0.665               | 265.2             |
| 0.10         | 0.206       | 0.670               | 225.0             |

**Table 2.** Relative MAP performance gain of using speech and metadata under different metadata quality conditions

## 6. CONCLUSION

We have presented an extension of the PSPL to incorporate spoken and text content information for document retrieval. The PSPL approach provides the flexibility to represent deterministic and stochastic document content information. This ability is used to propose a new relevance ranking framework that takes into account the different nature of the information sources available in the retrieval problem.

Moreover, experimental evidence supports the idea that exploiting the content of the spoken document does indeed provide a significant improvement in performance with respect to a scenario in which only the metadata is used for retrieval. This is an emblematic application scenario where the uncertain ASR information can be tolerated and positively used, providing significant performance improvement.

## 7. ACKNOWLEDGMENTS

We would like to thank Jim Glass and T J Hazen at MIT for providing the iCampus data. This work was supported by Microsoft Corporation.

## 8. REFERENCES

- [1] Ciprian Chelba and Alex Acero, "Position specific posterior lattices for indexing speech," in *ACL*, Ann Arbor, Michigan, June 2005.
- [2] Ciprian Chelba, Jorge Silva and Alex Acero, "Soft indexing of speech content for search in spoken documents," *ELSEVIER Computer Speech and Language*, in Press, 2006.
- [3] Sergey Brin and Lawrence Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107-117, 1998.
- [4] James Glass, T. J. Hazen, Lee Hetherington, and Chao Wang, "Analysis and processing of lecture audio data: Preliminary investigations," in *HLT-NAACL*, Boston, Massachusetts, pp. 9-12, May 2004.
- [5] Douglas W. Oard et al., "Building an information retrieval test collection for spontaneous conversational speech," in *SIGIR*, New York, pp. 41-48, 2004.