SpeakerLDA: Discovering Topics in Transcribed Multi-Speaker Audio Contents

Damiano Spina
damiano.spina@rmit.edu.au

Johanne R. Trippas
johanne.trippas@rmit.edu.au

Lawrence Cavedon
lawrence.cavedon@rmit.edu.au

Mark Sanderson
mark.sanderson@rmit.edu.au

School of Computer Science and Information Technology
RMIT University, Melbourne, Australia

ABSTRACT

Topic models such as Latent Dirichlet Allocation (LDA) [3] have been extensively used for characterizing text collections according to the topics discussed in documents. Organizing documents according to topic can be applied to different information access tasks such as document clustering, content-based recommendation or summarization. Spoken documents such as podcasts typically involve more than one speaker (e.g., meetings, interviews, chat shows or news with reporters). This paper presents a work-in-progress based on a variation of LDA that includes in the model the different speakers participating in conversational audio transcripts. Intuitively, each speaker has her own background knowledge which generates different topic and word distributions. We believe that informing a topic model with speaker segmentation (e.g., using existing speaker diarization techniques) may enhance discovery of topics in multi-speaker audio content.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]; H.3.3 [**Information Search and Retrieval**]

Keywords

Spoken Retrieval; Topic Modeling; Spoken Conversational Search

1. INTRODUCTION AND BACKGROUND

In the last decade, probabilistic generative models such as Latent Dirichlet Allocation (LDA) [3] have been proposed to characterize a document collection covering the topics underlying that collection. Thus, topic models have been adapted to incorporate information which is complementary to the lexical representation of documents such as time [1, 29] or author [24].

LDA and variations of it have been successfully applied in different information access tasks, including opinion mining and sentiment analysis [5, 31], topic detection [1, 23, 29], collaborative filtering [28] and word sense disambiguation [4]. Topic models have also been used to characterize audio content, particularly in the tasks of language model adaptation for speech recognition [10, 17] and topic segmentation [9, 12, 21].

We believe that topic models can be applied effectively to enhance access of spoken documents over an audio channel. There is an increasing consumption of spoken documents such as podcasts.[1] We hypothesize that incorporating intrinsic characteristics of such documents in topic models will improve the discovery of topics.

Not only is the use and production of spoken content growing, novel hands-free and eyes-free interfaces are being developed for the visually impaired[2] as well as for smart-phones and wearable accessories such as smart-watches. In such a context, clustering spoken documents according to topics could have application for presenting search results via a speech interface or to retrieve or cluster a set of podcasts on a given topic.

This paper presents a work in progress that proposes *SpeakerLDA*, an adaptation of LDA allowing for different *speakers* participating in multi-speaker spoken documents (e.g., interviews, chat shows or meetings). The intuition behind SpeakerLDA is that when the speaker is differentiated explicitly from the corresponding part of the audio, the model would be more accurate in discovering the underlying topics in the spoken document. For instance, the expert invited to a chat show may use terms specific to their field while the host may use more generic terms.

In this paper we describe the proposed topic model and identify possible applications that may benefit from using topic models. In this context, our long-term research question is as follows:

> *What is the impact in terms of effectiveness of adding speaker information into a topic model when compared to traditional LDA?*

A few assumptions are made in this work. Firstly, note that our aim is to discover topics at a document level (e.g., "Most probable topics for document d are z_1, z_2 and z_3") rather than trying to automatically segment spoken documents according to speakers or topics [9, 21]. Secondly, our topic model relies on a textual representation of spoken documents. For now, we assume that spoken documents are transcribed (either manually or by using Automatic Speech Recognition) and segmented according to speakers by using existent speaker diarization[3] techniques [25].

We next describe the proposed topic model SpeakerLDA. Some possible applications of topic models for accessing media content

[1] http://www.edisonresearch.com/podcast-share-of-ear/
[2] http://www.realthing.com.au/products/support-accessibility
[3] *Diarization* is the task of automatically segmenting a spoken audio into homogeneous partitions according to speaker identity.

SLAM'15, October 30, 2015, Brisbane, Australia.

ACM 978-1-4503-3749-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2802558.2814649 .

in audio channels are then presented in Section 3. We conclude and present future work in Section 4.

2. SpeakerLDA

Our proposed topic model consists of incorporating information about the speaker that participates in each segment of a spoken document. Differentiating between words *generated* by each speaker may enhance the model to obtain more representative topics.

Figure 1 shows the plate notation of LDA and SpeakerLDA. As a generative probabilistic model, LDA (Fig. 1(a)) tries to *explain* the observable variables of words w_w in a vocabulary N that occur in each document $d \in D$ by fitting the hidden variables θ_d and z_w, which are hyper-parametrized by α and β, respectively. The variable θ_d represents the multinomial distribution of k dimensions of topics in document d, while z_w represents the topic assigned to a given word w from a multinomial distribution over words.

SpeakerLDA (Fig. 1(b)) is an adaptation of the LDA model for documents that can be segmented according to the different speakers. Here, each speaker $s \in S$ has her own topic distribution θ_{ds}. That is, segments of a document associated with different speakers follow a different topic distribution.

Algorithm 1 sketches the generative process of SpeakerLDA.

Algorithm 1: Generative process for SpeakerLDA.

1: **for** document d in corpus D **do**
2: **for** speaker s in S_d **do**
3: Obtain pseudo-document d_s as the concatenated segments of d generated by speaker s
4: Choose $\theta_{ds} \sim$ Dirichlet(α)
5: **for** position w in d_s **do**
6: Choose a topic $z_w \sim$ Multinomial(θ_{ds})
7: Choose a word w_w from $p(w_w|z_w, \beta)$, a multinomial distribution over words conditioned on the topic z_w and the prior β
8: **end for**
9: **end for**
10: **end for**

Each document d can be seen as a composition of pseudo-documents d_s as the concatenated segments of d that have been generated by speaker s. For each speaker $s \in S_d$, the corresponding pseudo-document d_s is obtained (line 3). The same generative process as in LDA is then carried out. A *document-speaker*-specific topic distribution θ_{ds} is chosen by a Dirichlet distribution with prior α (line 4). θ_{ds} represents the topic distribution from which the speaker s chooses the topic used to generate each spoken word of document d. Therefore, from θ_{ds} the words in the different positions of d_s are generated (line 5). First, a topic indicator z_w is sampled from the multinomial distribution over topics (line 6); second, a word is chosen from a multinomial distribution over words conditioned on the topic z_w and the hyper-parameter β (line 7). As a result, the word-topic and document-topic distributions that generated the different pseudo-documents d_s are returned by the model.

In practice, SpeakerLDA is equivalent to splitting transcripts of spoken documents $d \in D$ according to their corresponding speakers $s \in S_d$ and running LDA over the collection of pseudo-documents d_s. In other words, it can be seen as adding an additional pre-processing step to the collection before running LDA. Therefore, the same sampling methods such as Gibbs sampling—and likewise, existent LDA implementations such as MALLET [18]—can be used to perform SpeakerLDA.

Note that SpeakerLDA returns a topic distribution for each speaker in the document. The topic distribution θ_d for a given document d could be obtained by combining the probabilities of each dimension of the vectors θ_{ds} for the speakers $s \in S_d$ via:

$$\theta_d = \Big\{ \sum_{s}^{S_d} \lambda_s \cdot \theta_{ds,i} \mid i \in 1..k \Big\} \quad (1)$$

where k is the total number of topics, $\theta_{ds,i}$ is the probability of the i-th topic in the distribution θ_{ds} for speaker s in document d, and λ_s represents the weight associated to speaker s, with $\sum_s^{S_d} \lambda_s = 1$. By default, this weight corresponds to the proportion of the document associated to the speaker. However, it can be used to explicitly give more importance to some speakers than others. For instance, it allows one to specify that the invited expert in a discussion has more impact on the final topic distribution of the document than the moderator of the discussion.

3. APPLICATIONS AND EVALUATION

We present tasks—related to accessing media on audio-channels—to which topic models could be applied. In particular, we describe the tasks of content-based recommendation for podcasts (Section 3.1) and search result clustering (Section 3.2). Section 3.3 then discusses an evaluation framework for addressing our research question.

3.1 Content-based Recommendation of Podcasts

Consider a user that consumes podcasts via an audio-only channel. Analogously to services such as YouTube, which suggests the next video to be played after the video that the user is watching, the system may recommend a *similar* audio to keep the user playing content. Here, the item to be recommended would be the closest document according to a given similarity function. One possible way to compute document similarity is to consider the topics assigned to the documents.

Comparing the topic distributions θ_d, θ'_d over two given documents d, d'—for instance, by computing Hellinger distance [19] or cosine similarity between the vectors inferred by a topic model—may enhance content-based recommendation or provide a complementary similarity function that could be combined with term overlap or recency.

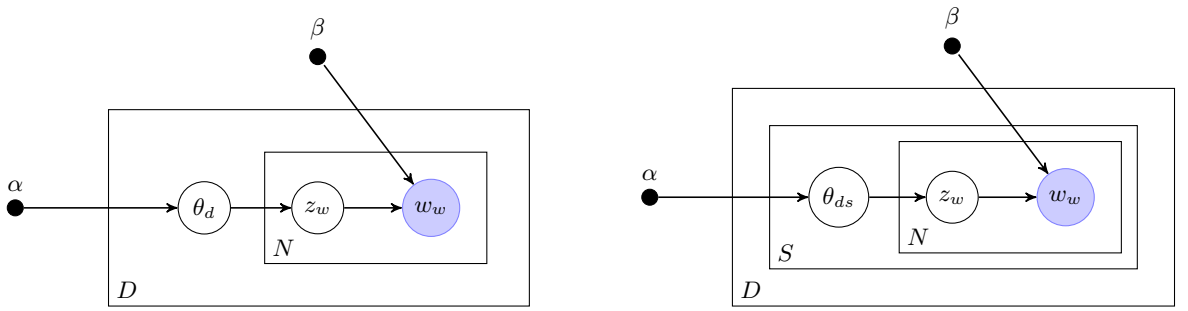
In the specific case of multi-speaker content, we believe that SpeakerLDA may obtain better topic representations, since it incorporates additional information over standard LDA.

3.2 Clustering for Search Result Presentation in Speech Interfaces

It is difficult to convey large amounts of information via speech without overloading a user's short-term memory [13, 26]. This is due to the nature of the spoken channel, where information can only be delivered sequentially. Therefore, presenting long lists of search results sequentially over a speech interface is unviable [22].

Since users often struggle with information overload [20], an alternative method to presenting results in a flat list is to cluster results [16]. Documents are grouped or clustered based on document similarities; thus, clusters maximize the coverage of the information space, allowing users to understand the important concepts as well as potential relationships between the search results [20].

Aggregating documents according to topic similarity may improve the application of current strategies for information presentation in dialog systems [8]. In particular, SpeakerLDA may out-



(a) Latent Dirichlet Allocation (LDA) (b) SpeakerLDA

Figure 1: Plate notation for LDA [3] and SpeakerLDA topic models.

perform existing topic models when applied to clustering search results in the context of multi-speaker podcast search.

Note that how to label topics is a collateral—and still open—problem to discovering topics [14] and little has been done on presenting topics in a speech interface.

3.3 Evaluation Framework

We now discuss factors that should be taken into account to measure the effectiveness of SpeakerLDA and its application to the tasks explained above. In terms of quantitative evaluation, topic models are typically evaluated by either computing intrinsic metrics of the model in an unseen set of documents [27, 29] or applied to external information access tasks such as information retrieval [30] or topic detection [23]. Measuring the human interpretability of topic distributions is itself also challenging [15, 27].

Comparing topic models to a manually annotated topic ground truth would give us better understanding of their effectiveness and their possible applications to clustering or recommendation tasks. Consider a collection of spoken documents which are manually assigned to a given set of topics. A topic distribution generated by a model can be translated to topic assignment by applying thresholds (e.g., document d belongs to topics in θ_d with probability greater than γ) or defining a similarity measure [19] to feed a clustering algorithm. Assigning documents to topics can be seen as clustering with an overlap, where each cluster corresponds to a different topic. In this context, metrics that quantify the precision and recall of clustering relationships can be applied [2].

A test collection must satisfy the following properties in order to be able to measure differences between our approach and existent topic models:

A. Each topic is discussed in two or more documents. If the ground truth does not contain topic relationships between documents, relationships captured by topic models cannot be evaluated.

B. Include spoken documents with two or more speakers. Note that if none of the documents include multiple speakers, performing SpeakerLDA for discovering topics would be equivalent to using LDA.

We now analyze the suitability of two annotated corpora available: the Fisher corpus [7] and AMI corpus [6].

The LDC Fisher corpus [7] comprises 5,850 transcribed telephone conversations between two subjects, each lasting up to 10 minutes. Participants typically do not know each other and, for each call, they are asked to speak on an assigned topic from a list of around forty topics such as “Pets”, “Family” or “Movies”.

Note that both speakers in the conversations included in the Fisher corpus talk about the same topic. Although conversations may stray to different off-topic issues [11], those are not explicitly annotated in the data. Having only one topic assigned to each document may not provide enough granularity to capture topic drift between the speakers. Because of this, the Fisher corpus it may not facilitate measuring differences between LDA and SpeakerLDA.

An alternative to the Fisher corpus is the Augmented Multi-party Interaction (AMI) corpus [6]. The AMI corpus comprises 100 hours of recorded meetings where participants play different roles in both real and elicited scenario-driven meetings. In addition to audio, the AMI corpus collects video, slides, and textual information. Audio is manually annotated with transcriptions, including speaker segmentation. Transcripts are segmented according to topics and subtopics. Since meetings typically contain more than two topic segments and assuming that documents share some topic/-subtopics in common (i.e., satisfies condition *A*), we believe that the AMI corpus is likely more suitable for studying the effectiveness of SpeakerLDA.

Further, speakers in the AMI corpus are assigned to play certain roles in a meeting, e.g., industrial designer, interface designer, marketing, or project manager. Conflating all speakers in the corpus who are playing the same role and then running SpeakerLDA may also improve discovery of the underlying topics in the meetings.

Nevertheless, having manual transcripts of the spoken documents will allow us to measure the effectiveness of running topic models over *perfect* segmented transcripts and thereby measure the robustness of different topic models against recognition errors.

4. CONCLUSIONS AND FUTURE WORK

We introduced SpeakerLDA, a topic model adapted to consider the speaker that generates each part of a multi-speaker document. Intuitively, each speaker has her own background knowledge, which generates different topic and word distributions. We believe that explicitly informing the model with speaker segmentation may improve the effectiveness of discovering topics in conversational audio content such as podcasts or meetings.

Immediate future work comprises developing an experimental setup that considers the AMI corpus as a test collection and compares the proposed approach with existing topic models. Moreover, we plan to study the impact of running topic models over noisy transcripts due to Automatic Speech Recognition and speaker diarization errors. Future work includes an extension of the model to incorporate cross-document *speaker identification* (i.e., multiple documents with the same speaker in common) which would enable,

for example, modeling the host and different guests that participate in a collection of podcast episodes.

5. ACKNOWLEDGMENTS

This research was partially supported by Australian Research Council Project LP130100563 and Real Thing Entertainment Pty Ltd. The authors wish to thank Ruey-Cheng Chen, who provided valuable feedback.

6. REFERENCES

[1] L. AlSumait, D. Barbará, and C. Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proceedings of ICDM'08*, pages 3–12. IEEE, 2008.

[2] E. Amigó, J. Gonzalo, and F. Verdejo. A general evaluation measure for document organization tasks. In *Proceedings of SIGIR'13*, pages 643–652, 2013.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] J. Boyd-Graber, D. M. Blei, and X. Zhu. A topic model for word sense disambiguation. In *Proceedings of EMNLP'07*, 2007.

[5] S. Brody and N. Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Proceedings of NAACL HLT 2010*, pages 804–812, 2010.

[6] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, et al. The AMI meeting corpus: A pre-announcement. In *Machine learning for multimodal interaction*, pages 28–39. Springer, 2006.

[7] C. Cieri, D. Miller, and K. Walker. The fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of LREC'04*, volume 4, pages 69–71, 2004.

[8] V. Demberg, A. Winterboer, and J. D. Moore. A strategy for information presentation in spoken dialog systems. *Computational Linguistics*, 37(3):489–539, 2011.

[9] L. Du, W. Buntine, and M. Johnson. Topic segmentation with a structured topic model. *Proceedings of NAACL HLT 2013*, pages 190–200, 2013.

[10] M. A. Haidar. *Language Modeling for Speech Recognition Incorporating Probabilistic Topic Models*. PhD thesis, Université du Québec, 2014.

[11] T. J. Hazen. Latent topic modeling for audio corpus summarization. In *INTERSPEECH'11*, pages 913–916, 2011.

[12] B.-J. P. Hsu and J. Glass. Style & topic language model adaptation using hmm-lda. In *Proceedings of EMNLP'06*, pages 373–381, 2006.

[13] J. Lai and N. Yankelovich. *Speech Interface Design*, pages 764–770. Elsevier, 2006.

[14] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin. Automatic labelling of topic models. In *Proceedings of ACL'11*, pages 1536–1545, 2011.

[15] J. H. Lau, D. Newman, and T. Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of ACL'14*, pages 530–539, 2014.

[16] A. V. Leouski and W. B. Croft. An evaluation of techniques for clustering search results. Technical report, DTIC Document, 2005.

[17] Y. Liu and F. Liu. Unsupervised language model adaptation via topic modeling based on named entity hypotheses. In *Proceedings of ICASSP 2008*, pages 4921–4924, 2008.

[18] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[19] D. Newman, S. Karimi, and L. Cavedon. External evaluation of topic models. In *Proceedings of ADCS'09*, 2009.

[20] H.-T. Pu. User evaluation of textual results clustering for web search. *Online Information Review*, 34(6):855–874, 2010.

[21] M. Riedl and C. Biemann. Text segmentation with topic models. *Journal for Language Technology and Computational Linguistics*, 27(1):47–69, 2012.

[22] N. G. Sahib, D. Al Thani, A. Tombros, and T. Stockman. Accessible information seeking. In *Proceedings of Digital Futures'12*, 2012.

[23] D. Spina, J. Carrillo-de-Albornoz, T. Martín, E. Amigó, J. Gonzalo, and F. Giner. UNED Online Reputation Monitoring Team at RepLab 2013. In *CLEF 2013 Eval. Labs and Workshop Online Working Notes*, 2013.

[24] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of KDD'04*, pages 306–315, 2004.

[25] S. E. Tranter, D. Reynolds, et al. An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1557–1565, 2006.

[26] M. Turunen, J. Hakulinen, N. Rajput, and A. A. Nanavati. *Evaluation of Mobile and Pervasive Speech Applications*, pages 219–262. John Wiley & Sons, Ltd, 2012.

[27] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of ICML'09*, pages 1105–1112, 2009.

[28] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of KDD'11*, pages 448–456. ACM, 2011.

[29] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of KDD'06*, 2006.

[30] X. Yi and J. Allan. Evaluating topic models for information retrieval. In *Proceedings of CIKM'08*, 2008.

[31] Z. Zhai, B. Liu, H. Xu, and P. Jia. Constrained LDA for grouping product features in opinion mining. In *Advances in Knowledge Discovery and Data Mining*, pages 448–459. Springer, 2011.