Term Clouds as Surrogates for User Generated Speech

Manos Tsagias
tsagias@science.uva.nl

Martha Larson
larson@science.uva.nl

Maarten de Rijke
mdr@science.uva.nl

ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam

ABSTRACT

User generated spoken audio remains a challenge for Automatic Speech Recognition (ASR) technology and content-based audio surrogates derived from ASR-transcripts must be error robust. An investigation of the use of term clouds as surrogates for podcasts demonstrates that ASR term clouds closely approximate term clouds derived from human-generated transcripts across a range of cloud sizes. A user study confirms the conclusion that ASR-clouds are viable surrogates for depicting the content of podcasts.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Algorithms, Measurement, Performance, Experimentation

1. INTRODUCTION

Podcasts, syndicated audio files published online, are growing in popularity and are available on the internet in ever increasing quantities [7, 8]. A challenge well-known from spoken audio in other forms is the design of audio surrogates, visual depictions that make it possible for users to assess the content of an audio recording at a glance [2]. Podcast surrogates are necessary so that aggregators or search engines can present users with easy-to-scan overviews of available content. Podcast episodes are typically represented by making use of the text information contained in their feeds, usually a title and description. However, such representations are dependent on the effort invested by the podcast publisher and may vary widely in level of detail. Content-based surrogates of podcasts generated from Automatic Speech Recognition (ASR) transcripts hold potential to provide searchers with richer, more complete representations of the topical content of a podcast episode.

Podcasts present a special challenge to speech recognition technology due to their heterogeneity. The fact that a significant portion of podcasts is user produced means that not only recording conditions and speaking styles can vary broadly, but also that there is no limit to the range of topics treated. Such conditions result in speech recognizer word error rates which hover around 40% [1, 5].

Motivated by work that has shown ASR-transcripts with high error rates can be used to generate keyphrase summaries [2] and snippet-like summaries [11], we investigate the question of whether term clouds automatically generated from ASR-transcripts are suitable podcast surrogates. These methods hold promise in the face of high word error rates characteristic of podcasts due to effects that collaborate to compensate for high error rates. The mitigating factors, summarized in [2], include the redundancy of speech, which ensures that words with high contribution to meaning occur repeatedly, raising the chance of correct recognition, and also the tendency of significant words to be mis-recognized less frequently.

Methods for generating and depicting tag clouds, visually weighted renditions of word sets, are becoming steadily more sophisticated [4, 9]. The popularity of tag clouds suggests that users find them helpful; they are frequently found on social web sites (e.g., del.icio.us, flickr) and blogs. Clouding techniques are also applied to terms occurring in a text [10] or document set [6], generating a semantic snapshot more appropriately designated a *term cloud*.

In this work, we assume as a baseline an optimal term cloud derived from reference transcripts of podcasts that were generated by humans. We introduce two measures, set-based and rank list-based, to compare reference transcript term clouds to term clouds derived from ASR-transcripts. The difference turns out to be quite modest, and a small-scale user survey confirms that differences between human-transcript clouds and ASR-transcript clouds have limited impact on the suitability of clouds to act as podcast surrogates.

2. EXPERIMENTS

2.1 From podcast to term cloud

For our experiments, we generated term clouds from 30 podcast episodes that we collected from three different U.S. English language podcasts: Security Now[1] (10 episodes, 9h 7m), British History 101[2] (10 episodes, 2h 7m), Music History Podcast[3] (10 episodes, 2h 48m). These podcasts contain a mixture of planned and unplanned speech. They were produced under reasonable recording conditions and contain short stretches of telephone speech and occasional background music. Not being recorded by professional speakers, these podcasts are considered user generated. Every episode includes a full transcript made available by the publisher. We estimated these transcripts to be at least 95% error free and we use them as reference transcripts to generate the baseline term clouds. ASR-transcripts are generated by the Nuance Dragon NaturallySpeaking 9 SDK Server Edition[4] speech recognition software used out of the box; no speaker enrollment or other adaptation was carried out. Term clouds are generated by counting how many times each word form occurs in the podcast episode, using either the

[1] http://www.grc.com/securitynow.htm
[2] http://bh101.wordpress.com
[3] http://www.musichistorypodcast.com
[4] http://www.nuance.com/audiomining/sdk

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

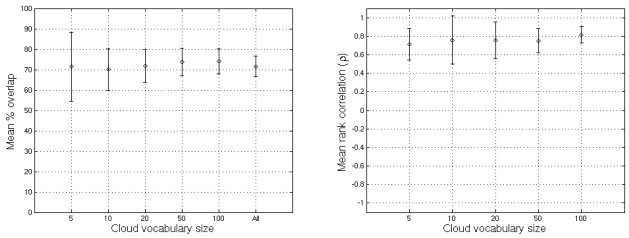

Figure 1: Word overlap and rank correlation between term clouds generated from reference and ASR-transcripts.

reference- or ASR-transcript. The top n most frequent terms are selected and visualized size-weighted according to relative frequency. Our cloud quality measures consider different term cloud sizes: $n \in N$, where $N = \{5, 10, 20, 50, 100\}$. For our user study we chose a typical value of $n = 50$. Stopwords are removed and lemmatization is applied to group words into normalized forms.

2.2 Measures of cloud quality

Our first measure of term cloud quality is set-based. We count the number of identical word forms in the vocabularies of term clouds of various sizes. Vocabulary overlap averaged over all podcast episodes is depicted in Figure 1(left) along with standard deviations. Moving from right to left, we see that both clouds containing all word forms in the vocabulary as well as clouds with vocabulary size $n = 100$ enjoy a peak overlap of 74% and that this percentage remains relatively stable as the size of the cloud decreases. This suggests that vocabularies overlap, and the relative frequencies of word forms remain comparable.

In order to gain a better understanding of the comparability of word weightings, we explore a second measure of term cloud quality that is rank-based. We calculate the Spearman's rank correlation coefficient between the reference term cloud vocabulary and the ASR term cloud vocabulary, both ranked by frequency. Figure 1(right) reports Spearman's rank correlation coefficient averaged over all episodes $\forall n \in N$. For clouds of size $n = 100$, the average correlation is 0.82, steadily declining as n decreases and hitting 0.63 for clouds of size $n = 5$. At $n = 5$, p-values were high on average ($\sim$0.5), expected due to the small size of the sample; at $n = 100$, p-values approach zero. These results suggests that relative weights are most comparable for larger tag clouds, but a basic similarity is also maintained in small-sized clouds.

2.3 User tests

In order to confirm that the similarity between speech-transcript clouds and human-transcript clouds demonstrated by the set-based and rank-based measures translates into utility for the searcher, we conducted a small-scale user study. The study investigated whether users performing a goal directed task requiring use of semantic information contained in the term clouds derived equal benefit from ASR-clouds and reference clouds. The goal of the task was to match the titles of 10 podcast episodes to the corresponding term clouds. Each participant performed the task once for each podcast (i.e., British History 101, Music History Podcast and Security Now), using the 10 episodes from each one. For each podcast, the participant was arbitrarily assigned either the ASR-cloud set or the reference cloud set. We decided to assess the utility of the term clouds by measuring task completion time (also used in [3]) and match accuracy. A group of 20 test subjects, mainly students, participated in the test. The task was carried out a total of 60 times, 20 times with title sets from each of the three podcasts. Of the 20 trials for each title set, the matching task was performed 10 times with ASR-clouds and 10 times with reference clouds.

On average, participants needed 3 minutes to match the reference clouds to the titles. When ASR-clouds were used, the average time needed to complete the matching task rose to approximately 3.5 minutes. This difference suggests a measurable but relatively small difference in utility between the ASR-clouds and the reference clouds. In particular, matches were performed correctly 92% of the time for ASR-clouds and 93% for reference clouds, supporting the conclusion that both types reflected the podcast content adequately.

3. CONCLUSION

We have found that term clouds derived from ASR-transcripts closely approximate term clouds derived from human-generated transcripts across a range of cloud sizes. Our user study confirms the conclusion that ASR-clouds are viable surrogates for depicting the content of podcasts. Several paths for future work open from here. User study participants remarked that prior knowledge of topic impacted their ability to use term clouds. This suggests adjustment of term clouds to user background or to their information need. Occasionally, phonetically similar tags appeared in the ASR-clouds (e.g., a British History 101 cloud contained both *knight* and *night*). Applying phonetic similarity to merge or cluster these terms might improve the term cloud utility. Finally, we are envisaging to make use of the feed-based metadata accompanying each podcast episode and create a mixed metadata/ASR cloud.

4. ACKNOWLEDGEMENTS

This research was supported by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104, and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 017.001.190, 640.001.501, 640.-002.501, STE-07-012.

5. REFERENCES

[1] W. Byrne et al. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Trans. Speech and Audio Processing*, 12(4):420–435, 2004.

[2] A. Désilets, B. de Bruijn, and J. Martin. Extracting keyphrases from spoken audio documents. In *Information Retrieval Techniques for Speech Applications*, pages 36–50, London, UK, 2002. Springer.

[3] M. Halvey and M. T. Keane. An assessment of tag presentation techniques. In *WWW '07*, pages 1313–1314, 2007.

[4] Y. Hassan-Montero and V. Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *International Conference on Multidisciplinary Information Sciences and Technologies*, 2006.

[5] K. Koumpis and S. Renals. Content-based access to spoken audio. *Signal Processing Magazine, IEEE*, 22(5):61–69, 2005.

[6] B. Y. L. Kuo and T. Hentrich and B. Good and M. D. Wilkinson. Tag clouds for summarizing web search results. In *WWW '07*, pages 1203–1204, 2007.

[7] K. Matthews. Research into podcasting technology including current and possible future uses. 2006. URL `http://mms.ecs.soton.ac.uk/2007/papers/32.pdf`.

[8] L. J. Patterson. The technology underlying podcasts. *Computer*, 39 (10):103–105, 2006.

[9] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen. Getting our head in the clouds: toward evaluation studies of tagclouds. In *SIGCHI '07*, pages 995–998, 2007.

[10] C. Shah and G. Marchionini. Discoverinfo: a tool for discovering information with relevance and novelty. In *SIGIR '07*, pages 902–902, 2007.

[11] X. Shou, M. Sanderson, and N. Tuffs. The relationship of word error rate to document ranking. In *AAAI Spring Symposium, Technical Report SS-03-08*, pages 28–33, 2003.