

# Investigating the Global Semantic Impact of Speech Recognition Error on Spoken Content Collections

Martha Larson<sup>1</sup>, Manos Tsagkias<sup>2</sup>, Jiyin He<sup>2</sup>, and Maarten de Rijke<sup>2</sup>

<sup>1</sup> Information and Communication Theory Group, EEMCS  
Delft University of Technology  
Mekelweg 4, 2628 CD Delft, The Netherlands  
[m.a.larson@tudelft.nl](mailto:m.a.larson@tudelft.nl)

<sup>2</sup> ISLA, University of Amsterdam  
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands  
{[e.tsagkias](mailto:e.tsagkias@uva.nl),[j.he](mailto:j.he@uva.nl)}@uva.nl, [mdr@science.uva.nl](mailto:mdr@science.uva.nl)

**Abstract.** Errors in speech recognition transcripts have a negative impact on effectiveness of content-based speech retrieval and present a particular challenge for collections containing conversational spoken content. We propose a Global Semantic Distortion (GSD) metric that measures the collection-wide impact of speech recognition error on spoken content retrieval in a query-independent manner. We deploy our metric to examine the effects of speech recognition substitution errors. First, we investigate frequent substitutions, cases in which the recognizer habitually mis-transcribes one word as another. Although habitual mistakes have a large global impact, the long tail of rare substitutions has a more damaging effect. Second, we investigate semantically similar substitutions, cases in which the word spoken and the word recognized do not diverge radically in meaning. Similar substitutions are shown to have slightly less global impact than semantically dissimilar substitutions.

## 1 Introduction

Sooner or later, a user searching with Google Audio Indexing<sup>3</sup> for videos related to Barack Obama will notice that the underlying technology upon occasion mishears “Barack” as “Broccoli.”<sup>4</sup> Are such speech recognizer substitution pairs merely amusing anecdotes, or could they hurt the overall effectiveness of a spoken content retrieval system? The literature on broadcast news retrieval reports that automatic speech recognition (ASR) errors do not stand in the way of satisfactory retrieval performance [6]. The redundancy of language, the tendency of a word to occur in a context containing similar words, compensates for speech recognition error. Research on spoken content retrieval has, however, moved into conversational domains such as interviews [2], lectures [3] and meetings [8]. These domains feature noisy background conditions and wide variation in speaking patterns [1, 4]. High levels of redundancy are no longer a given [7].

<sup>3</sup> <http://labs.google.com/gaudi>

<sup>4</sup> Search engine queried October 10, 2008

The less we can depend on redundancy the more we need to understand the impact of ASR error. In particular, we want to extend our understanding of error from the word/document level to the collection level. For example, if “Barack” is mis-recognized as “Broccoli,” a semantic connection is potentially introduced between two otherwise unrelated videos in the collection. If the impact of such errors is cumulative, the effects on spoken document retrieval could be substantial. This paper makes two contributions towards understanding the collection-wide impact of ASR error. First, we propose a Global Semantic Distortion (GSD) metric that provides an easily computable, query-independent measure of the global effect of error. Second, we show that for different ASR-error types, different effects arise. We examine high and low frequency ASR substitutions and ASR substitutions involving semantically related words.

Different types of ASR errors affect retrieval differently. Quality metrics alternative to word error rate have been proposed in the literature that better capture the suitability of ASR transcripts for retrieval [5, 9]. Our work extends these methods in that it goes beyond measuring the impact of error on a document in isolation to include capturing how error changes the semantic relationship of a given document with other documents in the collection. Our method is more general, since it evaluates the effects of error independently of particular queries. We attempt to understand ASR error within the framework of a larger research program aimed at reducing the influence of ASR error on retrieval. ASR-error compensating techniques familiar from the literature, such as domain adaptation, query/document expansion and use of ASR lattices [4], could potentially combine with methods based on our findings.

## 2 Global Semantic Distortion Metric

Our proposal for a Global Semantic Distortion (GSD) metric is based on a simple line of reasoning. ASR errors introduce semantically spurious words into document representations. If the words that co-occur in the context of a mis-recognized word are sufficiently redundant and reliable, the semantic impact of the ASR error at the document level will be minimal. Even if the semantic impact on the document is non-negligible, overall, the collection will suffer minimally, unless the ASR error shifts the meaning of the document such that it becomes semantically related to other documents in the collection.

We formulate our metric within the well-known vector space model. Proximity in the vector space is taken to reflect semantic similarity of documents. Our GSD metric takes the vector space defined by the human-generated reference documents to represent the original space, the undistorted semantic space of the collection. The GSD measures the level of semantic distortion that arises when the original space is transformed by replacing documents with noisy representations. First, we define the Semantic Distortion (SD) contributed by a single document as the fraction of its nearest neighbors in the original space that it loses when it is transformed by ASR error and shifts to the position of its noisy representation. More formally,

$$SD(doc_n, ASR_m) = 1 - \frac{\sum_i^{10} \sum_j^{10} \delta(PreShiftNeighbor_i, PostShiftNeighbor_j)}{10} \quad (1)$$

$$\delta(\text{neighbor}_i, \text{neighbor}_j) = \begin{cases} 1 & \text{if } \text{neighbor}_i = \text{neighbor}_j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where  $doc_n$  is the document,  $ASR_m$  is the set of ASR errors that transform the document,  $\delta$  is the similarity function in Eq. 2, and  $PreShiftNeighbor_i$  and  $PostShiftNeighbor_j$  are the nearest ten neighbors in the reference space before and after the document has been transformed by the introduction of ASR error. The *Global Semantic Distortion* is defined as the semantic distortion averaged over all documents in the collection,

$$GSD(C, ASR_m) = \frac{\sum_N SD(doc_n, ASR_m)}{N} \quad (3)$$

where  $C$  is the collection,  $N$  is the number of documents it contains and  $SD$  is the semantic distortion given by Eq. 1. The GSD requires no computationally intensive calculations; it only considers the nearest neighbors, set to 10 in exploratory experimentation. For the experiments presented below, we adopt the *tf·idf* weighting scheme and cosine similarity to determine the nearest neighbors, but the formulation of the GSD is general and admits alternate choices.

### 3 Data

We use the AMI Meeting Corpus (release 1.4) [8], which consists of 100 hours of multimodal data recorded from scenario-based meetings. Included in the corpus are automatic speech recognition transcripts and human-generated reference transcripts. This corpus is, to our knowledge, the largest publicly available corpus of conversational speech with reference transcripts in the form of full manual transcriptions. It is a suitable collection for our experiments since investigation of the global semantic impact of speech recognition error requires reliable reference transcripts for the complete spoken document collection. As is typical for conversational speech, the word error rate for the corpus ranges up to around 40% [8]. We use the speaker turn segmentation provided with the corpus to divide the data into documents. We use one half of the corpus as a development set to analyze patterns of ASR error and the other as a test set to measure the global impact of error. In order to eliminate the effects of interjections, we discard speaker turns consisting of less than 5 words, leaving us with a total test corpus of 3699 documents.

### 4 Experiments

In order to measure the impact of different types of substitution errors, we compute the GSD using documents with artificially controlled error levels. This approach gives us the ability to compare the relative GSD of different types of substitution error. We first remove all ASR errors (using oracle knowledge from the reference transcripts) to create document representations containing correctly recognized words only. Then we reintroduce ASR errors into the documents that were caused by a certain type of substitution error.

*Global impact of frequent substitutions:* In our first set of experiments, we use the proposed GSD metric to investigate the collection-wide semantic impact

**Table 1.** Global impact of ASR error due to frequent vs. infrequent substitutions as measured by the Global Semantic Distortion (GSD) metric

Condition	GSD
Correct words only	0.319
Correct words + errors due to frequent substitutions	0.348
Correct words + errors due to infrequent substitutions	0.363
Correct words + all errors	0.370

of ASR-error words introduced by *frequent substitutions* and *infrequent substitutions*. We are interested in discovering whether high-frequency (“habitual”) speech recognizer substitutions have a large global impact, since errors occurring frequently can be modeled more robustly than errors occurring infrequently. We investigate corpus statistics for word substitution pairs, pairs of words, e.g., (“cooperation,” corporation), the first spoken in the original audio, and the second substituted in by the speech recognizer. In total there are 67k substitution pairs, and we notice that they have a long-tail distribution. In particular, 36% of substitution pair instances are singularities (substitution pairs that occur only once in the development data). We divide the list into *frequent substitutions* (those with frequency  $\geq 4$ ) and *infrequent substitutions* (those with frequency  $\leq 3$ ). The cutoff is chosen to keep the total number of errors introduced in each condition balanced. The GSD of both conditions are reported in Table 1, along with the GSD caused when all ASR errors are retained. The GSD caused by representing documents using correctly recognized words is included as an upper bound. Although errors arising from frequent substitutions distort the space, the effects of errors arising from infrequent substitutions is higher and the difference is significant (Wilcoxon signed rank test, p-value  $< 2.2e-16$ ).

*Global impact of semantically similar substitutions:* In our second set of experiments, we use the proposed GSD metric to investigate substitution pairs in which the recognized word (i.e., the ASR error) has a very different meaning from the word originally spoken in the audio. In order to measure difference of meaning, we use the thesaurus-based ELKB application SimDist (<http://www.nzdl.org/ELKB>). This set of experiments finds motivation in the examination of select substitution pairs with high semantic similarity, such as (“bring,” keeping), (“worried,” worries) and (“bright,” right), and low semantic similarity, such as (“television,” innovation), (“tracking,” subtraction) and (“structure,” instruction). The contrast suggests that different levels of similarity have different potential for introducing semantic shift at the document level and ultimately at the collection level. We form a set of the substitution pairs that SimDist deems most similar and a set with an equal number of pairs that SimDist deems least similar and use these sets to build two contrasting experimental conditions. The

**Table 2.** Global impact of ASR error due to semantically similar vs. dissimilar substitutions as measured by the Global Semantic Distortion (GSD) metric

Condition	GSD
Correct words + errors due to dissimilar substitutions	0.362
Correct words + errors due to similar substitutions	0.360

results in Table 2 show that adding ASR errors arising due to the recognizer “mishearing” a word as a semantically dissimilar word has a small but significantly (Wilcoxon signed rank test,  $p$ -value  $< 4.258e-5$ ) greater impact.

## 5 Conclusion

This paper has made a contribution to the understanding of the effect of ASR error on spoken document retrieval by proposing a Global Semantic Distortion metric to capture the collection-wide impact of ASR error and by using this metric to analyze two types of ASR error. Our results show different types of ASR error exhibiting different levels of impact on the collection as a whole. In particular, error introduced by frequent ASR substitutions has a marked global effect. A more subtle difference in global effect is observed when the semantics of error words is taken into account. Errors involving a word semantically similar to the original spoken word are slightly less globally damaging than errors where the semantic dissimilarity is larger. Future work will involve refinement of the semantic distance metric used to determine semantically similar substitution pairs. Our ultimate goal is to use our understanding of the global impact of ASR error to improve spoken content classification and retrieval.

**Acknowledgments** This research was supported by the E.U. IST program of the 6th FP for RTD under project MultiMATCH contract IST-033104, and by the Netherlands Organization for Scientific Research (NWO) under project numbers 220-80-001, 017.001.190, 640.001.501, 640.002.501, 612.066.512, STE-07-012, 612.061.814, 612.061.815.

## Bibliography

- [1] J. Allan. Robust techniques for organizing and retrieving spoken documents. *EURASIP Journal on Applied Signal Processing*, 2003(1):103–114, 2003.
- [2] W. Byrnie et al. Automatic recognition of spontaneous speech for access to multi-lingual oral history archives. *IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing*, 12(4):420–435, 2004.
- [3] C. Chelba, J. Silva, and A. Acero. Soft indexing of speech content for search in spoken documents. *Computer Speech and Language*, 21(3):458–478, 2007.
- [4] F. M. G. de Jong, D. W. Oard, W. F. L. Heeren, and R. J. F. Ordelman. Access to recorded interviews: A research agenda. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, 1(1):3:1–3:27, June 2008.
- [5] J. Garofolo, E. Voorhees, C. Auzanne, and V. Stanford. Spoken document retrieval: 1998 evaluation and investigation of new metrics. In *Proceedings of the ESCA workshop: Accessing information in spoken audio*, pages 1–7, 1999.
- [6] J. Garofolo, C. Auzanne, and E. Voorhees. The TREC spoken document retrieval track: A success story. In *RIAO 2000*, 2000.
- [7] G. Jones, K. Zhang, E. Newman, and A. Lam-Adesina. Examining the contributions of automatic speech transcriptions and metadata sources for searching spontaneous conversational speech. In *ACM SIGIR SCS 2007 Workshop*, 2007.
- [8] S. Renals, T. Hain, and H. Bourlard. Recognition and understanding of meetings: The AMI and AMIDA projects. In *ASRU 2007*, 2007.
- [9] L. van der Werff and W. Heeren. Evaluating ASR output for information retrieval. In *ACM SIGIR SCS 2007 Workshop*, 2007.