International Journal of Innovative Computing, Information and Control
Volume **6**, Number **3(B)**, March **2010**

ICIC International ©2010 ISSN 1349-4198
pp. **1525–1536**

SIGNAL BOOSTING FOR ROBUST DATA FUSION IN SPEECH RETRIEVAL

Dan Wu[1] and Daqing He[2,*]

[1]School of Information Management
Wuhan University
Luojia Hill, Wuhan 430072, P. R. China
woodan@whu.edu.cn

[2]School of Information Sciences
University of Pittsburgh
135 North Bellefield Avenue, Pittsburgh, PA 15260, USA
*Corresponding author: dah44@pitt.edu

Received January 2009; revised May 2009

Abstract. *Data fusion has been widely used in Web retrieval to combine evidence from multiple sources. However, its effectiveness is still unknown in the speech retrieval tasks where retrieval results in diverse quality can be generated from the multiple representations of speech documents. This paper describes an investigation of the signal boosting techniques for data fusion in speech retrieval, whose goals are to enhance the signals from relevant documents and to reduce the effect of noise coming from low quality retrieval results. Our retrieval experiments performed on the MALACH speech document collection demonstrated that CombMNZ, the most widely used data fusion method, cannot handle the noise in our speech retrieval tasks. We, therefore, developed and examined two new fusion methods called WCombMNZ and WCombMWW. Our experiments showed that the new methods can significantly outperform CombMNZ in the speech retrievals on the MALACH collection, and they have great potential in more general data fusion tasks too.*
Keywords: Data fusion, Speech retrieval, Signal boosting, CombMNZ, MALACH

1. **Introduction.** With the vast amount of Web information in various formats or media, the search for relevant documents for a given topic often requires various improvement methods, which include fuzzy reasoning [20,21], personalization [12] and data fusion [35]. Data fusion refers to the techniques of combining evidence from multiple sources [17]. In the literature, two possible scenarios of data fusion have been commonly discussed. First, the integration can happen at the query side, where terms from multiple sources are combined to generate a single query. Examples include Belkin and his colleagues' study on combining Boolean queries from multiple users [5], and query expansion based on relevance feedback [13,26], which has been explored recently in Text REtrieval Conference (TREC) Robust and HARD tracks [2,31]. The second scenario of integrating multiple pieces of information happens at the result side. In this scenario, results retrieved for one topic (query) but from different search engines, they are combined to generate one single outcome. Web meta-search engines often take this route [24].

Searching on spontaneous speeches has become common along with the development of Web technology and especially with the further development of automatic speech recognition (ASR) technology [23]. However, spontaneous speech is considerably more challenging for the ASR techniques than commonly used broadcast news. For example, the MALACH collection [22] is widely used in speech retrieval evaluation tasks. It contains 7800 manually constructed segments from 300 interviews of Holocaust survivors. Because