**General Comments:**

This paper is the first comprehensive analysis of search term performance on a problem of scale. Keyword search is an increasingly important problem. The genesis for this work was an evaluation conducted by NIST in 2006. The technology analyzed in this paper has been in operational use, particularly in the intelligence community, for some time, on extremely large amounts of operational data. The paper is an attempt to bring some basic science to problems that have been consistently observed by operational systems. The authors can assure you that the observations in this paper are informed by extensive operational use. Unfortunately, this information is classified, and no such comparable resources exist in the public domain on the unclassified side. This work is intended to begin a movement in this direction, and to influence an upcoming keyword search evaluation.

The reviewers’ comments about the lack of basic science in the paper must be tempered by the observation that the human language technology field typically focuses on applications of technology. Algorithms are most often borrowed from other fields, such as statistics, and rarely invented with the field. Most journal papers consist of adaptations of these algorithms to specific problems of interest to the community. This paper falls in the latter category. We make no claims to the development of novel algorithms. It is study of the correlates between search terms and keyword search performance. It is the first such study on a critical mass of data. Keyword search technology, similar to speech recognition technology, needs vast amounts of data if one is to attempt to draw conclusions about the trends in performance for a wide variety of words.

The importance of this work is in the exploration of what features correlate with keyword search term performance. There are a vast number of linguistic features that could be considered. We have, in fact, considered over 150 combinations of various features. It would make sense to explore some of the specific combinations in greater detail. For example, reviewer no. 1 commented about the lack of discussion about BPC. However, in reality, most of these linguistic features carry about the same amount of information and individually are not great predictors of performance. Hence, we pursued an aggregate approach.

A significant contribution is to reduce this vast space to something more manageable. For example, it was well known before this work that duration was a primary correlate. However, we have done a good job of exploring this relationship in more detail. The goals of this work are very clearly stated in the introduction to the paper. This topic has not been previously explored in the literature, though error analysis has been performed on a much more limited and anecdotal basis previously. This work will form the basis for a more extended evaluation being conducted this year. We hope as data from these new studies become available we can extend the models and gain new insight into the problem.

Below we address the specific concerns of the reviewers:

**Reviewer #1:**

**1. Authors have not given any insights about the explored features and models.**

**2.The paper is not technically strong due to lack of analysis of results**

**3.The paper is not scientific, in the sense that there is no logic or justification in using various features and models.**

**6. In the present draft, there is no technical contribution. Authors have used existing features and models blindly, and got some numbers. There is no interpretation regarding the selection of features and models. There is no analysis and interpretation about the obtained results.**

These four concerns are essentially restatements of the same issue.

The revised paper contains slightly more discussion of the explored features and models. It must be pointed out that the entire section on acoustic analysis resulted from an analysis of the poor performance obtained from the phonetic distance approach. We originally approached the work only from a phonetic basis. Unfortunately, those features had limited value. That triggered a far deeper analysis of these factors, which in turn gave rise to a major piece of this paper. So the claim that there was no detailed analysis provided seems to be due to a lack of understanding of the relationship between these two approaches and the motivations for the acoustic approaches.

The underlying, or root causes, of errors is discussed in the context of the degree to which acoustic matching is influenced by phonetic constraints. We have added some additional comments to better underscore the analysis presented in this paper. However, it must be emphasized that the key contribution of the paper is an exploration of which features are most highly correlated. That in itself represents an analysis of search term performance. The underlying causes of why particular features are good or bad is discussed broadly in terms of some acoustic and phonetic issues, but it is difficult to draw such conclusions for particular features in a statistically meaningful manner. Anecdotal evidence only goes so far in this type of analysis.

Some analysis is provided throughout the paper whenever possible. For example, on page 10 we discuss the role the parameter “count” plays in the results. We also discuss duration on page 5. We discuss and analyze the construction of the data sets in Table 2 and 3.

We have added more detailed explanations of the experiments throughout Section IV. Hopefully this will not put us over the page limits.

**4.Details about existing literature are missing.**

**5.There is no comparison with the state of the art in this area.**

As mentioned previously, this is the first comprehensive study of this type of problem that we are aware of. That is what makes the paper unique. We have accurately cited all the previous work related to the specific systems and data presented in this study. In fact, the authors worked closely with the providers of these data and systems. This paper was not meant to be an extensive or exhaustive tutorial on keyword search.

Further, all prediction algorithms used are referenced, even though the details of these algorithms is outside the scope of this work.

If the reviewer feels we left out particular references, we would appreciate more specific guidance, as long as these are relevant to the study presented.

**7. In Table-1 sounds are categorized into various groups based on their manner of articulation. Later there is no discussion on, how these groups effect the recognition performance.**

Space limitations, of course, impact the extent to which every experiment can be discussed. There are over 1,000 experiments using over 150 features that were conducted to bring the study to this point, and it is difficult to give justice to each one of these. The entries labeled SFS in the various tables represent combinations of various features. If necessary, we can go into greater detail of the specific subset of features selected. We decided not to do this since the relative differences between these combinations are fairly small. We focused on the explicit features that had the greatest impact on the problem.

**Reviewer #2:**

**First, this paper is really screaming for some sort of user evaluation. I'm not talking about a controlled user study, I am talking about some preliminary qualitative work as to whether speech-naive users actually improve their search performance using the tool.**

**There is a web interface, but it is mentioned only in passing. I would like to see more discussion of this. If the authors can provide, my evaluation of the paper will go up considerably.**

User evaluations are difficult to conduct. We have provided a web-based demo and have observed some naïve and expert users use of this demo. However, the demo is connected to a relatively limited database of indexed audio provided by Microsoft. There is no doubt that the feedback provided by our prediction algorithm is useful. However, stating this in a scientifically meaningful manner is difficult because user experiences with keyword search are complex. Despite significant advances in the technology in the past decade, user perceptions of the technology, particularly in the intelligence community, vary greatly.

At our university lab, we do not have the resources to conduct such studies. It should be noted that the 2006 evaluation, upon which this work was based, also did not conduct user studies. Few HLT papers present such results in a meaningful manner, and traditionally focus more on objective measures of performance.

One of the authors has extensive experience with the use of this technology in intelligence applications. As mentioned in the paper, dealing with users’ frustrations with keyword search formed the basis for this work. Unfortunately, this information is classified and cannot be discussed in the open literature. However, let me emphasize that keyword search technology has been extensively evaluated by the authors in very real operational conditions. The author has worked directly with analysts on strategies for selecting keywords, and those experiences have informed the research in this paper.

For example, it was long conjectured that search terms with long durations were better. Unfortunately, many search terms used in operational settings tend to be short. It has been a complicated process to train users on the shortcomings of these approaches. There are strong interactions between search terms, the domain of interest, and the acoustic confusability of these terms. Anecdotal evidence of a few users searching the limited audio archives available on the Internet (such as the site our web-based demo links to) have only limited scientific value.

Again, our lab does not have the resources to conduct such evaluations on the unclassified side, and that really goes beyond the scope of this paper. This paper is not meant to be a comprehensive evaluation of search term technology. It is meant to explore a small piece of this overall puzzle – what phonetic and acoustic features correlate with search term performance.

**Second, the approach considers only the term being searched, not the corpus or the task. Thus it somewhat confounds standard text-based search techniques, which users are probably already using. For example, based on Figure 2, searching for "shopping" may be a \*bad\* idea, if the term is found frequently, as it will give too many results and these will have to be searched tediously. So there is also some tension here. Perhaps the work would be best thought of as part of a total search strategy, that might also consider frequency of terms in the corpus as well as similar sounds terms in the corpus which might be candidate hits (see, e.g. the work of Vemuri).**

We acknowledge that data is always a problem in these kinds of studies. When we began this study, we attempted to acquire evaluation data from all the major systems presented at the 2006 NIST evaluation. We were able to capture most of it. However, the database used in that study turned out to be too small and suffered from some of the problems the reviewer alluded to. Therefore, we worked with several of the system providers to get more data, and that is why we used Fisher. Even Fisher is moderate for this kind of analysis. If a search term appears less than 50 times in a corpus, it is very hard to draw conclusions about its behavior. On the other hand, analysts want to be able to search for terms that are not commonly occurring words.

This is why, for example, we explored underlying linguistic features such as N-grams of phonemes. Unfortunately, even N-grams of phonemes occur somewhat sparsely, even in Fisher. This motivated us to look at features such as N-grams of BPC. Such features occur in multiple words and hence we get more stable estimates of their overall performance. This is a detail that might not be obvious to the average reader.

We believe we have been very careful to emphasize the differences between text-based searches and audio-based searches. We agree that for a given false alarm rate, if a term occurs frequently, the number of actual hits returned will be greater than for a less frequently occurring term. Unfortunately, for the less frequently occurring term, the recognizer will also have fewer examples to train on. Hence, there is a complex relationship between acoustic decoding and what amounts to language model probabilities.

Intelligence analysts, or people who use keyword search technology frequently in their day-to-day jobs, do understand these tradeoffs at an intuitive level. They tend not to use generic terms that occur frequently. Since they are searching vast amounts of data, they must use terms that are generally occurring infrequently and of high value. However, within this range of terms, there is significant latitude in which specific terms they can pick. A good example might be choosing the difference between a word like “vote” and a word like “democracy” or “government”. The technology provided in this paper gives them some insight into the difference between such terms.

**Summary Comments:**

**All in all I think the paper is solid and a small contribution. I do question the overall utility of the approach in a real search task and would like to see some evidence that it does have value there. But it is also a short paper and I would be happy to see it published if other reviewers are positive.**

We would like to thank the reviewers for their thoughtful and insightful feedback. We have done our best to address each of their concerns. We hope our responses have adequately addressed your concerns and we can proceed with publication.

We want to emphasize that this is the first study of its type to be published in a major journal as far as we know. The timing is important since there are significant evaluations of keyword search planned for the coming years. It makes sense to establish this as an area of science that can support and motivate additional analyses. As more data becomes available, the scientific quality of these studies can be refined.