Enabling Accent Resilient Speech based Information Retrieval

Koushik Sinha
Hewlett Packard Labs, India
sinha_kou@yahoo.com

Geetha Manjunath
Hewlett Packard Labs, India
geetha.manjunath@hp.com

Raveesh R. Sharma
Hewlett Packard Labs, India
raveesh.sharma@hp.com

Viswanath Gangavaram
Hewlett Packard Labs, India
viswanath.gangavaram@hp.com

Pooja A
Hewlett Packard Labs, India
pooja.a@hp.com

Deepak R. Murugaian
Hewlett Packard Labs, India
deepak-raj.murugaian@hp.com

ABSTRACT

Voice interfaces to browsers and mobile applications are becoming popular as typing with touch screens is cumbersome. The main issue of practical speech based interfaces is how to overcome speech recognition errors. This problem is more severe when the users are non-native speakers of English due to differences in pronunciations. In this paper, we describe a novel, intelligent speech interface design approach for IR tasks that is significantly robust to accent variations. Our solution uses phonemic similarity based word spreading and semantic information based filtering to boost the accuracy of any ASR. We evaluated our solution with Google Voice as the ASR for a web question-answering system developed in-house and the results are very encouraging.

Categories and Subject Descriptors

H3.3 [**Information Search and Retrieval**]: [speech interface]

General Terms

Design, Algorithms.

Keywords

Speech based IR, QA, semantic feedback, phonemic spreading.

1. INTRODUCTION

Speech based interfaces are experiencing growing popularity in diverse application domains such as mobile information retrieval (e.g., Apple Siri, AT&T Speak4It). While *automatic speech recognition* (ASR) systems with small closed vocabulary have acceptable recognition accuracy, recognition of natural language speech input still needs research. An unfortunate side-effect of tailoring an ASR for a specific set of words is that a fixed vocabulary may be impractical when using an ASR for information retrieval (IR).

Among the vast research on speech recognition, one promising approach that is being preferred in recent times is to combine an open-domain ASR with a post-error correction module to create an acceptable application-level performance [1]. Taking this approach to the next level, we believe that enabling a tighter semantic coupling between an IR system and the ASR (treated as a black box) would considerably help in solving the problem. While popular ASRs, like Google Voice, are typically good, they are unable to provide high recognition accuracy for accented English. As some of these ASRs are offered from the cloud, it is also not possible to fine tune them for different users. Our intelligent, error-correcting speech-based IR system is resilient to wide variations in pronunciations caused by presence of noise and differences in accents.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

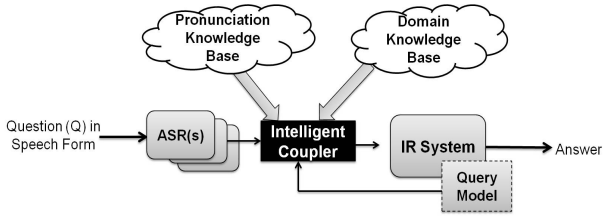

Figure 1: Proposed intelligent speech interface based IR system

2. PROPOSED SOLUTION

For most IR tasks, it is sufficient to correctly recognize the key terms in a spoken query. If we can exploit the semantic relationship between the terms in a natural language query to correct the key terms, we would gain in accuracy. Thus, by modeling the key aspects of inputs expected by the IR task and by exposing the right interfaces in the IR system for validating them, one can develop an intelligent feedback module for an ASR to considerably improve the end-to-end accuracy of IR. We illustrate this using a web question-answering (QA) system as a representative IR system.

Figure 1 depicts the high level architecture of our solution. The *intelligent coupler* that sits between the ASR system (consisting of one or more ASRs) and the IR system utilizes a *pronunciation knowledge base*, a *domain knowledge base* and a semantic feedback on the ASR's text output from a *query model* block within the IR system. For the rest of the paper, we refer to the input speech data as Q and the corresponding text output from the ASR as Q^*.

The pronunciation knowledge base (KB) is a mapping of words to their corresponding *phoneme sequence(s)* that characterize the way(s) a word is pronounced. A *phoneme is a speech unit that encapsulates a unique utterance or sound*. The domain knowledge base (KB) contains domain specific information that can aid error correction of Q^* by limiting the search space. The *query model* represents the underlying structure (e.g., semantics) of typical queries submitted to the IR system.

Figure 2 shows the details of the *intelligent coupling block* for QA (called ICB). The ICB consists of four major components: i) phoneme sequence generator (PSG) module, ii) spreading dictionaries, iii) sound similarity module (SSM) and, iv) question model builder and evaluator (QMBE) module. The PSG takes as input

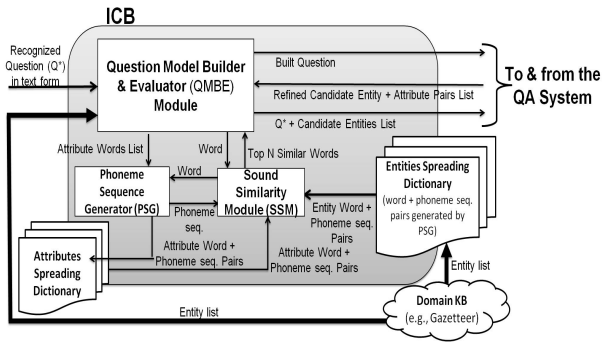

Figure 2: Interaction between the various components

a word w and generate a phoneme sequence ϕ_w representing an approximate pronunciation of w using a graphone model [2]. A *spreading dictionary* S consists of $(w(S), \phi_w(S))$ pairs generated using PSG. Given a (w, ϕ_w) pair, SSM creates a ranked list of words with phonemic similarity to $\phi_w(S)$'s. For this, four features are first computed between a ϕ_w and every $\phi_w(S)$: i) their longest common subsequence (LCS), ii) their mean similarity score (MSS) as per the algorithm in [3], iii) common prefix length and, iv) common suffix length. Next, a ranked list R of $(w(S), \phi_w(S))$ pairs is created using SVM^{rank} [4]. The semantic coherence between the various sentence elements present in Q^* is evaluated using the query/question model of the QA system. For simplicity, we assume a question model consisting of (a) an optional interrogative phrase, (b) an entity name and (c) one or more attributes associated with that entity name.

The QMBE first generates a ranked list R_1 of k_1 ($k_1 > 0$) candidate entities by calling the SSM module with the *entities spreading dictionary*. The *entities spreading dictionary* is created offline using a set of entities extracted from some domain KB such as a gazetteer or the Wikipedia. Next, it uses a set of validation primitives (such as isEntity() and hasAttributes()) associated with the query model to generate a filtered list of candidate entity-attribute pairs from R_1. This list is used to generate the *attributes spreading dictionary* which is then utilized to generate a ranked list R_2 of k_2 ($k_2 > 0$) of candidate attribute terms phonemically closest to the potential attribute word(s) in Q^*. Finally, R_1 and R_2 are combined to form a ranked list R_{12} of k model questions consisting of (entity, attribute) pairs, using the *normRank* rank combining function [5].

3. RESULTS

For performance evaluation, we formed an *evaluation dataset* of 220 questions matching our simple question model. The entity names and their corresponding attributes were extracted from the key-value pairs in the Wiki Infoboxes corresponding to various topics (e.g., places, people, etc.). A sample question in the dataset is "*Who is the president of India*".

We measure the performance of a speech-driven QA system in terms of correctly recognizing a given input question Q and the individual words in Q. For this, we have used two metrics, namely sentence-level accuracy (SLA) and word-level accuracy (WLA) defined as follows. Let N be the total number of questions and N^* the number of questions from N in which all three elements of the question model were correctly recognized. Let W be the total number of words present in the N questions. Let W^* be the total number of correctly recognized words from W (including repetitions). Then, $SLA = N^*/N$ and $WLA = W^*/W$.

We evaluated three popular ASRs - PocketSphinx 0.6, Microsoft

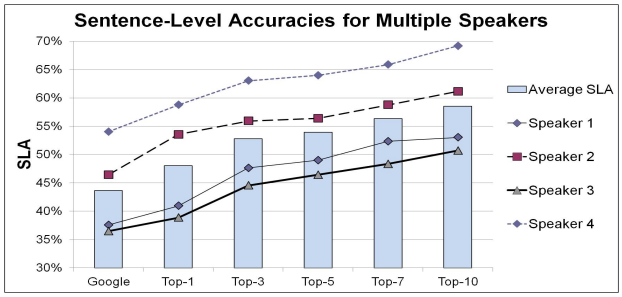

Figure 3: Performance comparison for different accents

SAPI 5.3 and Google Voice on a subset of 55 questions from the dataset with a single speaker. While the WLA results were 71.7%, 65.16% and 87.38% respectively, the SLA results were 22.8%, 29.82% and 52.63%. As seen from these results, sentence recognition is more difficult than recognizing individual words, for all ASRs. Given that Google Voice performed best, we compared the performance of our solution (using Google Voice as the frontend ASR) against Google Voice alone on the complete 220 questions dataset using four speakers with different accents. On WLA comparison, the mean WLA for Google Voice and the ICB (with $|R| = 1$) computed for the four speakers were 82.58% and 85.60% respectively. For SLA comparison, the ICB creates a ranked list of top-k model questions conforming to the question model. We consider our solution to correctly recognize a question Q if any of these k model questions has all three sentence elements of Q.

The results of this experiment are depicted in Figure 3. Top-1, Top-3, Top-5, Top-7 and Top-10 refer to the corresponding values of top-k model questions built by our proposed ICB. In Figure 3, the mean SLA for Google Voice was 43.64% when averaged over the SLA values for all the four speakers. In contrast, the mean SLA values for the ICB were 48.03%, 52.79%, 53.95%, 56.33% and 58.52% for k = 1, 3, 5, 7 and 10. The mean SLA values are represented as columns against the corresponding x-axis labels.

4. CONCLUSION

We have presented a new approach to designing speech based IR systems robust to accent variations. On our evaluation data set, our solution shows an improvement by 10.31% in SLA for $k = 5$ and by 3.02% in WLA, over Google Voice. Since our technique uses semantic relationship across terms in a sentence, this comparatively smaller improvement in WLA was expected. As future work, we intend to explore more complex query models with a richer dataset.

5. REFERENCES

[1] M. Jeong and G. G. Lee, "Improving speech recognition and understanding using error-corrective reranking," *ACM Trans. on Asian Lang. Inf. Processing*, vol. 7(1), pp. 2:1–2:26, 2008.

[2] Sequitur G2P, `http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html`, 2011.

[3] B. Hixon, et al., "Phonemic similarity metrics to compare pronunciation methods," *Proc. Interspeech*, 2011.

[4] SVM^{rank}, `http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`, 2011.

[5] Y. Wang, X. -G. Qi, B. D. Davison, "Standing on the shoulders of giants: ranking by combining multiple sources," *Tech. Rep. LU-CSE-07-011*, Lehigh University, USA.