A Mobile Voice Search Client for European Portuguese

Pedro B. Pascoal[1,2] Daan Baldewijns[1] Fernando Miguel Pinto[1]
Manuel Ribeiro[1] José Santos[3] Miguel Sales Dias[1,4]
[1]Microsoft Language and Development Center, Lisbon, Portugal
[2]INESC-ID/IST/Technical University of Lisbon
Microsoft, UK
[4]ISCTE - University Institute of Lisbon/ADETTI-IUL, Lisbon, Portugal
{t-pedrop, v-daanb, a-fpinto, t-manrib, jcsantos, miguel.dias}@microsoft.com

Summary
The generalization of speech technology availability notably in simple mobility scenarios, such as searching the Internet using the user´s voice or "voice search", has increased user awareness of the usefulness of this human-computer interaction modality. In this paper, we present a mobile voice search client that accepts a generic spoken query and presents the search engine's result page using European Portuguese.

Keywords
Voice search, Mobile application, Windows Phone, Portuguese speech corpora.

1. INTRODUCTION

With the dissemination of speech technology, namely in mobile devices such as smartphones, the demand for voice enabled features has significantly increased. Today´s users require not only the ability of using voice instructions in their mobility experience, but also of searching generic topics in a "hands-free" fashion [1].

Usability evaluation studies, such as reported in Teixeira et al. [2], suggest that speech is the easiest and most natural modality of human-computer interaction. Speech is also the preferred modality when interacting with mobile devices, especially when faced with "on-the-go" situations, like for example when driving that make it temporary difficult to all, the use of other interaction modalities like touch. Taking these requirements into consideration, we have developed a Voice Search client for Windows Phone devices that accepts a generic spoken query using European Portuguese Language that is then recognized by a cloud-based system using the Microsoft Public Speech platform. The resulting text string is then feed into Microsoft´s Bing search technology, thus producing the required Bing result that was requested using the user´s voice. In this paper, we describe our language model (LM) training process, since it is one of the key components of a Speech Recognition technology. A LM which assigns probabilities to sequence of *n* words and predicts the next word in a speech sentence.

2. TEXT CORPUS FOR LANGUAGEM MODELING

A Language Model should be calculated from text data that is similar to the domain where it is applied to. In our case, we would require large amounts of correctly transcribed spoken web queries to build a consistent language model. Since there was no data available that represents spoken web search queries for European Portuguese, we needed to sample users' written queries from a search engine. Therefore, we sampled five months of actual queries from Microsoft *Bing* logs for the Portuguese market to create the text corpus for language model training. It has been proven that mixing data sources from different search query logs can improve the quality of Voice Search systems [3].

3. SPEECH CORPORA FOR SPEECH RECOGNITION TESTING

The most frequent web queries of the training set where identified and where later used to build our test corpus of spoken search queries. Using an internal platform for the collection of speech data through mobile devices, we collected almost 2,5 hours (about 29,785 word tokens) of pure read speech from 83 speakers, that were eliciting such most frequent web queries. Note that written queries sampled from search engines often contain incorrect spellings of words. In order to use these queries, we required a pre-processing step in which we clean the training data by applying a set of rules to it.

4. LANGUAGE MODEL TRAINING

To train a language model, we need to normalize the data in a pre-processing step followed by the training or language model building phase.

4.1 Pre-Processing

It is the goal of this pre-processing stage to normalize the data so that the same queries and the same notions are expressed in the same way. For instance, the conversion of queries to lower-case, the correction of misspell words

or the removal of unknown character, are enforced at this stage.

4.2 Language Model Building

The training stage uses the normalized training corpus and builds the language model. The first step counts the unigrams (occurrence of each word in the set) in the normalized training set and builds a vocabulary file. Each word in such file has its phonetic transcription available in an existing Lexicon, or such transcription is calculated by an available Letter To Sound algorithm. The availability of a phonetic transcription of each word in the vocabulary file is required for speech recognition to be successful. With the pre-processed vocabulary list, we are able to compute bigram and trigram counts. Those counts are used to build a language model, in which probabilities are assigned to each n-gram sequence.

5. VOICE SEARCH SYSTEM

In order to use the trained language models in a real-world scenario, we have developed a mobile client application that allows a user to submit general web queries to a search engine on a mobile device, using the European Portuguese language. The system can be divided in a client-side, with graphical user interfaces, and a cloud-based server-side, with the speech recognition services

5.1 Graphical User Interface

Since the mobile application was developed for Windows Phone, both the user interface, interactive experience and feedback were made similar to the user interface natively available in Windows Phone for other languages (Figure 1). This way, the usage of our application should require little to none learning when used by current Windows Phone users. The spoken query is captured by the mobile device and sent to the speech recognition services. Once received the recognized query, the client performs a web search on Bing for the received query, displaying the results much like when doing a text web search.

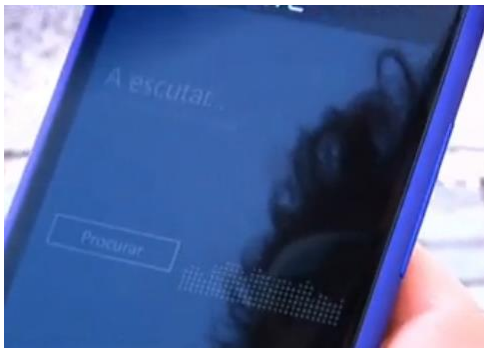

Figure 1 - User performing a voice search query in our system.

5.2 Speech recognition services

The Speech-Recognition services are deployed in a public the cloud backend infrastructure, allowing for higher scalability and availability of the platform. The backend provides a REST-based API, built on ASP.Net Web API technology, running on the Windows Azure cloud platform. This gives the possibility of using specially designed language models. Since these language models have significant processing and memory requirements to be used, we leveraged the use of the cloud to minimize loading and recognition time, by adopting such strategies as keeping a pool of pre-instantiated instances for more resource intensive language models, which can be re-used between requests. Resorting to this architecture allows us to offer this set of services, not only for this application, but also for future applications, with minimal engineering effort. Additionally, the server is able to collect and store (with due user consent) users' spoken queries (utterances), from multiple clients, providing a feedback loop where the collected text data can be used to re-train our language models and the retrieved spoken data, can be also used to re-train our acoustic models, using untranscribed data techniques.

6. CONCLUSION AND FUTURE WORK

In this work, we presented a Voice Search client-server system that accepts generic spoken queries in Portuguese. As we acquire more data and learn users' intentions, we can update the system with better and more reliable components, since the queries performed by users can be used to train more robust language and acoustic models. Future improvements of the system on language modelling, will use larger amounts of training data and perform improvements on the normalization rules to further increase its precision. Expanding this feature to more languages available in the Microsoft Public Speech Platform (http://www.microsoft.com/pt-pt/mldc/downloads.aspx), is also in our plans.

7. ACKNOWLEDGEMENTS

This work has been carried out in the scope of the QREN 11495 World Search project, co-funded by Microsoft, the Portuguese Government, and the European Structural Funds for Portugal (FEDER) through COMPETE (Operational Program for Competitiveness Factors), as part of the National Strategic Reference Framework (QREN), the national program of incentives for Portuguese businesses and industry. The authors are indebted to the speakers involved in the data collection effort.

8. REFERENCES

[1] Maryam Kamvar and Doug Beeferman. Say what? why users choose to speak their web queries. In Interspeech, 2010.

[2] Teixeira, A., Braga, D., Coelho, L., Fonseca, J., Alvarelhão, J., Martín, I., Queirós, A., Rocha, N., Calado, A., Dias, M.: Speech as the basic interface for assistive technology. In: Proc. International Conference on Software Development for Enhancing Accessibility and Fighting Info-Exclusion. Porto Salvo (2009)

[3] Li, X., Nguyen, P., Zweig, G., Bohus, D.: Leveraging Multiple Query Logs to Improve Language Models for Spoken Query Recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 3713-3716. (2009)