# PhD Topic Proposals



## Practical information

- The PhD activities listed in this document can be executed at the Nuance locations in Merelbeke, Belgium; Aachen, Germany or Ulm, Germany. The choice of location can be agreed upon between Nuance, the candidate student and the University.
- The PhD thesis will be done in cooperation with the Dialogue Systems Group at Ulm University. The candidate will work as an integral member of the research team at Nuance Communications he or she will be associated with.
- The PhD thesis will be supervised by Ulm University.
- Application material an inquiries should be sent to wolfgang.minker@uni-ulm.de

## Spoken query and information retrieval in large databases: General problem description

Several use cases for speech recognition relate to retrieval of information from a structured database, such as retrieving destinations from a map database, retrieving items from a music database or looking up contact information from address books. Since the arrival of natural language interfaces in the mobile world such as Siri and S-voice, users expect systems to not impose restrictions on how to phrase their request, but be capable to understand the user intent in a given context, regardless of how it is phrased.

The databases of interest consist of a large number of entries. Navigation databases contain information such as city name, street name and house numbers. For points-of-interest, the actual POI name, brand names and categories may be given as well. In a spoken query to a large database, the user enters potentially partial information into the system. If sufficient information is entered, the query corresponds to a specific entry of the database. If the utterance matches a (potentially large) set of the database, tagged information elements are returned to the application for further disambiguation steps. Information given by the user can also be incorrect, out of domain, or refer to an item that is not stored in the database, which needs to be detected and returned to the application as well.

For traditional information retrieval, measures such as TF/IDF are typically used. The ranking of a document, which is part of a larger set of documents, queried by a set of search words is determined on the frequency of the words in the documents relative to the entire set. The more often a search word appears in a particular document and the less likely it appears in the overall set, the higher the score for this document of the word is in the query. The Jaccard similarity coefficient is another statistic to measure the similarity of sample sets. Other scoring mechanisms interpret the problem as an NLU task, and compare the returned interpretation with a reference interpretation.

Let us assume for example the query “Berlin”. A valid response of the application is that the city of Berlin was meant and that there are more than ten thousand addresses and POI entries connected with it. However “Berlin” is also used in other contexts. There may be cafes and hotels named “Berlin”, so one may wonder if this answer really is the perfect response.

Things get more complicated when “Berlin” is part of an utterance. “Bring me to Berlin” would support the standard hypothesis that the city of Berlin was meant. The utterance “Take me to the [Hotel] Berlin” may imply that returning the hotel “Berlin” is more appropriate, if there is one near the current

location of the user or near the destination. The true intent can thus be hidden in the carrier phrase rather than the phrasing of the slot content itself.

If one assumes several database types can be queried at the same time, the utterance “Berlin” may even refer to the band of that name in a database of music items, and have nothing to do with a geographical location.

From these examples, it can be seen that determining the reference interpretation is a non-trivial, often ambiguous and context sensitive problem.

Any technology being developed needs to take deployment cost into consideration. This refers to the CPU and memory resource cost, as well as the application development and localization cost of any of the underlying modeling techniques. Rule based systems require manual work to develop the rules with the drawback that the quality is highly dependent on the proficiency of the developer. Data driven systems move the problem to the efficient collection of statistically significant amounts of data, and can potentially create a cost bottle neck in the annotation stage of the process. Therefore, a lot of attention needs to go to automation of processes, including automating the validation of the system’s performance.

# Topic1: Quality Assessment Method and Tooling

The main goal of this topic is to derive and implement information theoretic measures to measure the accuracy of a speech database retrieval system.

The intent is to provide and implement appropriate quality metrics for spoken queries in large databases, correlated with user perception, applied in a production system without requiring expensive manual tagging. The metrics shall be developed and tested using empirical evidence on real data. Furthermore tools will need to be developed to support tagging and evaluation of the system on datasets of recorded utterances, starting from and becoming part of an existing eco-system of development tools. These will then be integrated in a monitoring system that tracks the progress of the technology in terms of accuracy as well as resource consumption.

The quality assessment method should be independent of the technology used to solve the database query problem, enabling impartial comparison of technological options or solutions. It should also include information about statistical significance of the results, both in absolute terms as well as in the context of a comparative study of several systems.

# Topic2: Resource Efficient Natural Language Database Query by Voice

The main goal of this topic is to derive and implement a resource (CPU, RAM, ROM) efficient way to retrieve the most likely item set of a large data base corresponding to a spoken query, and measure the accuracy of such system with the metrics developed in 0. The proposed solutions need to be deployable on embedded systems.

The goal is to develop a search engine that uses information theoretic criteria to search for the most appropriate answer to a voice query. An important aspect in speech recognition, unlike text queries, is to cope with uncertainties and acoustic decoding errors. Therefore the information theoretic measures have to be combined with acoustic scores as well as language model scores, in a principled way rather than heuristic. It should be noted that queries can include natural language phrasings of the intent (action), though the emphasis of this topic is to retrieve the database object (mention) rather than detecting the intent itself.

The solution will build on existing components and tools as much as possible, with a preference to extend their capabilities. New paradigms can be introduced under the condition that they will significantly outperform legacy implementations in accuracy and/or functionality.

# Topic3: Natural Language Database Query by Voice in Context

This topic addresses the general problem of finding the most relevant documents or database items by spoken query utterances in large databases in a given context. The baseline to this work is a system which performs this task purely by acoustical evidence under the restrictions of a given database. The main problem found in the baseline approach lies in the ambiguity of speech utterances towards potential document candidates, which oftentimes cannot be resolved by pure acoustic evidence.

The main goal of this item is to improve the document retrieval by using additional information beyond the acoustics and database content. Such information can be derived from semantic information extraction stemming from the carrier phrase of the utterance itself, as well as external contextual information provided by the application. Contextual information includes, but is not limited to, dialogue state, location of the vehicle, knowledge of the navigation route, state of the vehicle, user history and user preferences. The system will need to work in the context of an existing eco-system of semantic classifiers, possibly suggesting additions or modifications for optimal accuracy.

The overall system can be comprised of several loosely coupled subsystems, each performing a given task. Several databases of different or similar domains can be active, together with other speech related components. The proposed solution will be able to function in such context, and support an arbitration function to decide on a final ordered list of hypotheses with associated probabilities to be presented to the application.