# Bachelorarbeit - Better accuracy of automatic lecture transcriptions by using context information from slide contents

Jonathan Werner

## Contents



# 1 Introduction {-} Scannability is crucial for academic research: you

have to be able to quickly evaluate the usefulness of a given resource by skimming the content and looking for the parts that are specifically relevant to the task at hand.

The medium in which those resources are available is very centered on textual representation. Spoken content, hereinafter called **speech media** (audio- or audiovisual media that mainly consists of spoken language) **doesn't make it possible to scan its contents.** You are "stabbing in the dark" when looking for something specific in a medium like this and have to consume it like a linear narrative.

This means that although lectures and conference talks are a central element to science they are much more challenging and tedious to use for research work.

Being able to a) efficiently **search** and b) look at the **temporal distribution of important keywords** in a visually dense way would elevate the usefulness of speech media in the scientific context immensely.

One approach to accomplish those goals is utilizing Automatic Speech Recognition (ASR) to transcribe speech to text and also get timing information for the recognized words. This makes it possible to derive information about the density of given words at a given point of time in the talk, which in turn allows to compute **word occurence density maxima**. This opens up possibilities for **compact visual representation** of the interesting keywords, thus allowing the user to **scan**.

The main challenge when using ASR for this task is the recognition accuracy of technical terms. Most of them are not included in the language models that are available as those are broad and generic so as to optimize for accuracy over a wide topic spectrum. But when they are not included into the language model they have no chance to be correctly recognized at all.

So the usefulness of applying ASR with a generic language model to the problem is very small, as the intesection of interesting keywords with those technical terms that can not be recognized is very big.

The central goal of this thesis is to explore an approach to overcome this problem. This approach consists of using words from lecture slides or other notes to **generate a lecture-specific language model**. This is then **interpolated** with a generic language model and being compared to the 'baseline' accuracy of the generic model.

## Structure of this thesis

The structure of this thesis is laid out as follows:

(1) **Research questions**

I will state the research questions.

(2) **Scientific Background**

(a) I will start by giving an overview over the state of the art of ASR and the most prevalent approaches.

(b) I will explain the *concepts* which are fundamental for the understanding of speech recognition.

(c) I will then examine the *scientific work* that has been done on applying ASR to the problem of lectures transcriptions.

(d) Finally i will summarize the *metrics* that have been used to assess the quality of the improvements in different approaches.

(3) **Motivation**

Here i will provide a motivation why it is necessary to improve on the baseline performance of ASR in our context.

I will talk about keywords and technical terms and why they are not being detected and how that diminishes the usefulness of ASR for the purposes of scannability.

(4) **Test data**

I will use the openly available *Open Yale Courses*, which provide a diverse selection of university lectures with the added bonus of having quality transcriptions and course notes or slides available.

I will present the chosen courses, their selection criteria and discuss the range of types of lecture material.

(5) **The LM-Interpolation approach**

(a) **Technical basis**
I will introduce the open source speech recognition framework *Sphinx 4*. This is the software that is used for performing the actual recognition.

(b) **Process overview**
I will first give a overview of the design and architecture of our approach.

(c) **Implementation**
I will then describe the technical implementation by which the lecture material is compiled into a specialized language model and recognition is performed using a *interpolated* language model.

(6) **Analysis**

(a) **Methods**
I will discuss how to analyze the results and develop metrics that assess how well the given goals are met with our approach.

(b) **Analysis**
I will then perform quantitative analysis on our test dataset with the metrics we developed before.

(c) **Discussion, Finding and Conclusions**
I will discuss the findings.

I will then draw conclusions from the quantitative analysis concerning the viability of our approach.

(1) **Improvements, Open Ends**

I will discuss possible improvements and open ends that were out of the scope of this thesis but would be interesting to further exploration.

(2) **Summary**

I will end by summarizing the goals, the proposed approach, the design and implementation, the analysis and the results.

# 2 Research questions

The central research questions i want to investigate in this thesis can be formulated like the following:

(1) When we want to run ASR on speech media, especially university lectures, what is the advantage of using an approach that consists of creating a lecture-specific language model and interpolating it with a generic language model, given that we are interested in improving the recognition accuracy of *interesting keywords* for the sake of searchability and scannability?

(2) What metric is useful for quantifying this advantage?

A secondary question is:

> How can we *use* the results from our approach to provide graphical *interfaces* for improving the users ability to search and scan the given speech medium?

The exploration of this question will not be the center of this thesis, but it will provide practical motivation for the results that the exploration.

# 3 Background

## 3.1 The field of Automatic Speech Recognition

Automatic Speech Recognition (ASR) can be defined as the process by which a computer maps an acoustic speech signal to text [1].

Rabiner and Juang [2] date the first research on ASR back to the early 1950s, when Bell Labs built a system for single-speaker digit recognition. Since then the field has seen three major approaches, which Marquard [3] summarizes as follows:

1. The *acoustic-phonetic approach* aimed to identify features of speech such as vowels directly through their acoustic properties, and from there build up words based on their constituent phonetic elements.

2. The *statistical pattern-recognition approach* measures features of the acoustic signal, and compares these to existing patterns established from a range of reference sources to produce similarity scores which may be used to establish the best match.

3. *Artificial intelligence (AI) approaches* have been used to integrate different types of knowledge sources (such as acoustic, lexical, syntactic, semantic and pragmatic knowledge) to influence the output from a pattern-recognition system to select the most likely match.

The most prevalent approach today is the *statistical pattern-recognition approach*, as it produces results with much higher accuracy compared to the acoustic-phonetic approach. The use of Hidden Markov Models (HMM) has been playing a key role in this approach, as it allows recognizers to use a statistical model of a given pattern rather than a fixed representation.

In the last years there has been a resurgence of AI approaches, specifically *deep learning approaches* [4]. The ASR paradigm we will use for this thesis will be limited to the former, however.

## 3.2 Dimensions of speech recognition

There are three dimensions which serve to classify different applications of speech recognition [1] [3]:

(1) **Dependent vs. independent**. Dependent recognition systems are developed to be used by one speaker, whereas independent systems are developed to be used by *any* speaker of a particular type, i.e North-American speakers. **Adaptive** systems lie between those poles, they are able to adapt to a particular speaker through training.

(2) **Small vs. large vocabulary**. Small vocabularies contain only up to a few hundred words and might be modeled by an explicit grammar, whereas large vocabularies contain tens of thousands of words so as to be able to model general purpose spoken language over a variety of domains.

(3) **Continuous vs. isolated speech**. Isolated speech consists of single words that are spoken with pauses in between them, whereas continuous speech consists of words that are spoken in a connected way. Continuous

speech is significantly more difficult to recognize, as it is a) more difficult to find the start and end of words and b) the pronunciation of words changes in relation to their surrounding words.

# Bibliography

[1] "Comp.speech frequently asked questions." http://www.speech.cs.cmu.edu/comp.speech/Section6/Q6.1.html.

[2] L. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," 1993.

[3] S. Marquard, "Improving searchability of automatically transcribed lectures through dynamic language modelling," 2012.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and others, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.