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

Jonathan Werner

Contents

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) **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.

(4) **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.

(5) **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.

1 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.

2 Scientific Background

2.1 The field of ASR, prevalent approaches

Stephen Marquard distinguishes three approaches of speech recognition [1]:

(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.

Of these approaches, the statistical pattern-recognition approach produced significantly better accuracy than the acoustic-phonetic approach, and is now the dominant paradigm for speech recognition, augmented by various AI approaches. A key element in pattern recognition is the use of Hidden Markov Models (HMMs), which enables recognizers to use a statistical model of a pattern rather than a fixed template.

2.2 Fundamental ASR concepts

2.2.1 Recognition scope

Speech recognition applications can be broadly characterised in three ways: speaker dependent or independent, small or large vocabulary, and isolated or connected recognition.

Speaker-dependent systems are designed to recognize speech from one person, and typically involve a training exercise where the speaker records sample sentences to enable the recognizer to adapt to the speaker's voice. Speaker-independent systems are designed to recognize speech from a wide range of people without prior interaction between the speakers and the recognition system.

Small vocabulary systems are those where only a small set of words is required to be recognized (for example fewer than 100), and permissible word sequences may be constrained through a prescriptive grammar. Large vocabulary systems are those designed to recognize the wide range of words encountered in natural speech (for example up to 60,000 words).

Finally, isolated recognition systems are intended to recognize a discrete word or phrase, typically as an action prompt in an interaction between person and system, whereas connected recognition systems are intended to recognize continuous words and sentences following each other without interruption.

Three possible applications and their characteristics are shown in Table 2-1. Application Speaker Vocabulary Duration Dictation Dependent Large Connected Command and control system Independent Small Isolated Lecture transcripts Independent Large Connected Table 2-1: Characteristics of some common speech recognition applications

The subfield relevant to the creation of automatic transcripts from lecture speech is thus characterised as speaker-independent (SI) large vocabulary connected (or continuous) speech recognition (LVCSR).

2.2.2 Acoustic and Recognition systems following the dominant statistical pattern-recognition paradigm

1. A set of phonemes
2. A phonetic dictionary
3. An acoustic model
4. A language model

A phoneme is a unit of sound making up an utterance. The most general representation of phonemes is that provided by the International Phonetic Alphabet (IPA), which includes orthography for phonemes found in all oral languages [14].

However, for speech recognition applications, ASCII representations of phonemes are more practical. A widely used ASCII set is the Arpabet (Table 2-2), created by the Advanced Research Projects Agency (ARPA) to represent sounds in General American English [15]. The Arpabet comprises 39 phonemes each represented by one or two letters with optional stress markers represented by 0, 1 or 2.

An acoustic model associates features from the sound signal with phonemes. As the pronunciation of an individual phoneme is affected by co-articulation effects (how sounds are pronounced differently when voiced together), many systems model phoneme triples, i.e. a phoneme in context of the phonemes preceding and following it. As the exact pronunciation and sound of a phoneme may vary widely, even from a single speaker, acoustic models reflect probabilities that a set of acoustic features may represent a particular phoneme (or set of phonemes).

Acoustic models are trained from a speech corpus consisting of audio recordings matched with a transcription. The transcription typically contains time-alignment information to the word- or phoneme level. Speaker-independent models are trained with audio from a wide range of speakers (for example with a mix of male and female speakers and regional accents). Speaker dependent models may be trained from a single speaker, or more commonly, created by adapting a speaker independent model to a given speaker.

However, acoustic models alone are insufficient to achieve acceptable levels of accuracy, as can be illustrated by the challenges of disambiguating between homonyms and similar-sounding phrases such as “wreck a nice beach” and “recognize speech”. Linguistic context is thus an additional and indispensable resource in generating plausible recognition hypotheses.

The dominant approach to language modelling is the n-gram language model (LM). Such language models are trained from a text corpus, and give the probability that a given word will appear in a text following the (n-1) preceding words. Smoothing techniques are often applied to the initial model to adjust the probabilities to compensate for the fact that less frequent words which have not been seen in the training text may also occur.

For example, Table 2-4 shows the probabilities for words which might follow "your economic" in a trigram (3-word) language model in ARPA format:

```
-2.0429 YOUR ECONOMIC ADVISERS
-1.2870 YOUR ECONOMIC FUTURE
-2.0429 YOUR ECONOMIC GROWTH
-1.7585 YOUR ECONOMIC POLICIES
-1.7585 YOUR ECONOMIC POLICY
-1.1613 YOUR ECONOMIC PROGRAM
-2.0429 YOUR ECONOMIC PROGRAMS
-1.5947 YOUR ECONOMIC PROPOSALS
-2.0429 YOUR ECONOMIC REFORM
-2.0429 YOUR ECONOMIC REFORMS
-1.3695 YOUR ECONOMIC TEAM
```

In this example, where the recognizer is assessing which hypothesis is most likely for a word following "your economic", the language model would favour "program" rather than "programs", and "team" over the homonym "teem". However, the model would give no advantage to the recognizer in distinguishing between singular and plural forms of "reform" and "policy" as they are equally likely in the model. Language models are used in a range of natural language processing applications, including spell-checkers (to suggest the most likely correction for a misspelt word) alanguage models

2.2.3 Speech corpora

Training acoustic and language models require appropriate corpora. Notable corpora used in speech recognition research have included:

The TIMIT corpus of American English speech (1986), which consists of a set of sentences each read by a range of different speakers [15].

The Wall Street Journal (WSJ) corpus (1992), derived largely from text from the Wall Street Journal newspaper from 1987-1989 read aloud by a number of speakers [17].

The HUB4 English Broadcast News Speech corpus (1996/7), generated from transcriptions of news programmes broadcast in the United States on CNN, CSPAN and NPR [11], [18].

The Translanguage English Database (TED) corpus (2002), created from lectures given by a range of speakers at the Eurospeech '93 conference [19]. The above examples have each been carefully curated to serve research purposes, and are derived from specific genres or application domains. Models trained from such corpora may be less effective when applied to different contexts. For example, acoustic models trained by American English speakers may be less effective for recognizing 12speech from other parts of the world, and language models trained

on broadcast news may be less effective when applied to a different genre, such as poetry.

2.3 Work on applying ASR to lecture transcription

Over the last decade, a number of research groups and projects have undertaken systematic work in applying speech recognition to lectures, progressively investigating multiple techniques and approaches. Major programmes include:

work by the Spoken Language Systems Group in the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT [5]

work by Cosmin Munteanu and colleagues in the Computer Science Department at the University of Toronto [20]

the Science and Technology Agency Priority Program in Japan, "Spontaneous Speech: Corpus and Processing Technology", supporting work particularly at

Kyoto University and the Tokyo Institute of Technology [21]

the Liberated Learning Project [22]

the Net4voice project under the EU Lifelong Learning Programme [23]

the Computers In the Human Interaction Loop (CHIL) project under the EU FP6 [24].

The starting point of speech recognition research for lectures is usually recognition systems developed for earlier applications. These include broadcast news and meeting transcription systems, or speaker-dependent systems such as those used for dictation. Speaker independent systems are typically trained with widely available speech and language corpora, such as those described in 2.2.3.

As initial results in applying the recognition systems and accompanying acoustic and language models to lecture speech usually produced poor results characterised by high error rates, much of the related research effort has focused on improving the effectiveness of speech recognition for lectures through different types of generalization and specialization of earlier systems and approaches.

Generalization approaches have examined ways of accounting for the larger vocabulary, including specialized terms, and greater variation in delivery style characteristic of spoken lectures. Specialization approaches have looked at features specific to many lectures, such as the use of presentation slides, and using these attributes to "know more" about the content of the lecture and thus improve recognition accuracy and usefulness.

A further class of research starts by accepting the imperfect nature of automatically generated transcripts, and examines how to involve users in improving transcript accuracy and where possible use correction feedback to further improve subsequent automated recognition tasks.

2.4 Modelling the form, style and content of lectures

The form and linguistic style of lectures present both challenges and opportunities for ASR systems.

For example, Yamazaki et al note the high level of spontaneity in lectures, which are characterized by "strong coarticulation effects, non-grammatical constructions, hesitations, repetitions, and filled pauses" [25]. Glass et al note a colloquial style dramatically different to that in textbooks, characterized by poor planning at the sentence level and higher structural levels [26]. Lectures additionally exhibit a high number of content-specific words which may not occur in a general speech corpus. Spoken and written forms of language may diverge differently in different languages; for example, Akita and Kawahara note significant linguistic differences between spoken and written Japanese [27].

These variations have presented recognition difficulties, and a range of strategies have been explored to compensate.

Structural features in the genre have been observed and exploited to improve recognition performance. Such features include rhetorical markers, the use of presentation slides, a close correspondence between speech and slides or textbook, and affinity between the content of related lectures and between lectures and associated material available on the web.

Most models of the lecture for ASR systems assume a single speaker engaged in a monologue in a single language, accounting for students or the audience only so far as they constitute a potential source of noise. Birnholtz notes a lack of "systematic study of face-to-face behavior" in the research related to webcasting systems, focusing particularly on audience interactivity and how turn-taking ("changes in floor control") is dynamically negotiated [28].

14Although a sub-field of speech recognition known as speaker diarization is devoted to identifying multiple speakers in audio (typically in the context of meetings or conferences) [29], the potential requirement for ASR systems to transcribe not only the speech of the lecturer but also that of people asking questions or interjecting in a lecture is largely unexplored.

2.5 Acoustic model adaptation Acoustic models derived from the broadcast and news genres may be a poor fit for lecture recordings, and thus a class of research has focused on how to adapt acoustic models to more accurately reflect the characteristics of lecture speech. Adaptation strategies which have shown some success include accounting for non- linguistic speech phenomena ("filler" sounds) [30], dynamically adjusting the model to account for speaking rate [31], unsupervised adaptation to account for new speakers [32] and using discriminatively trained models for language identification and multilingual speech recognition [33].

2.6 Language model adaptation Researchers have investigated strategies for generating and adapting the language model (LM) to improve recognition accuracy for lectures, on the assumption that a model which closely reflects the

context of the utterances is likely to outperform a more generic language model. Adaptations have been investigated for three levels of context:

at the macro level, for all lectures, treating spoken lectures as a genre with distinct characteristics

at the meso level, for a single lecture, taking advantage of prior knowledge about the lecture topic or speaker

at the micro level, for a part of a lecture, using knowledge about segments or transitions within a lecture.

Many adaptation strategies make use of some prior knowledge or parallel media. This could include information about the topic or knowledge domain of the lecture, a textbook or instructional materials related to the course or the lecture presentation slides. Use of such information may provide specific improvements at the expense of the generality of the technique (for example, not all lectures may be accompanied by slides). Kato et al investigated the use of a topic-independent LM, created by creating a large corpus of text from lecture transcripts and panel discussions, with topic-specific keywords removed [34]. The model is then adapted to specific lectures by using the preprint paper of the lecture to be delivered (when available).

Willett et al propose two iterative methods of unsupervised adaptation [35] [36]. Both methods show improvements in accuracy up to the second iteration of application. A first method identifies texts from a large corpus which are considered close to the first-pass recognition text by using a Term Frequency – Inverse Document Frequency (TF-IDF) measure, and uses the selected texts to adapt the LM. TF-IDF is a weighting factor which assigns a score to the importance of the word based on its occurrence in 15the document (term frequency) but adjusted to avoid words which are common across all documents (such as "a" and "the") from dominating the score.

A second method uses a minimum discriminant estimation (MDE) algorithm to adapt the LM, following the thesis that "seeing a word uttered at some place within the speech increases the likelihood of an additional appearance". MDE is a technique for adapting a language model to more closely match the distribution of words seen in the recognized text, while minimizing the variation from original to adapted model, using a measure of distortion (or discrimination information) known as the Kullback- Leibler distance. [37] Nanjo and Kawahara report similar work, and further explore adaptations to the lexicon and LM to account for variant pronunciations [38].

The use of lecture slides for adapting the LM has been explored by several research groups. Yamazaki et al note that a "a strong correlation can be observed between slides and speech" and explore first adapting the LM with all text found in the slides, then dynamically adapting the LM for the speech corresponding to a particular slide [25]. Munteanu et al pursue an unsupervised approach using keywords found in slides as query terms for a web search. The documents found in the search are then used to adapt the LM [39].

Kawahara et al investigate three approaches to adapting the LM, viz. global topic adaptation using Probabilistic Latent Semantic Analysis (PLSA), adaptation with web text derived from keyword queries and dynamic local slide-by-slide adaptation using a contextual cache model. They conclude that the PLSA and cache models are robust and effective, and give better accuracy than web text collection because of a better orientation to topic words [40]. Latent Semantic Analysis is an approach to document comparison and retrieval which relies on a numeric analysis of word frequency and proximity.

Akita and Kawahara propose a statistical transformation model for adapting a pronunciation model and LM from a text corpus primarily reflecting written language to one more suited for recognizing spoken language [27]. While n-gram language models are the dominant paradigm in ASR systems, they offer a relatively coarse model of language context. Newer research is exploring more accurate statistical representations of "deep context", for example accounting for connections between related but widely separated words and phrases [41].

2.3.1 Metrics

Bibliography

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