T-61.184 : Speech Recognition and Language Modeling : From Theory to Practice
Project Groups / Descriptions
Fall 2004
Helsinki University of Technology

Project #1: Music Recognition

Jukka Parviainen (parvi@james.hut.fi)
Ville Turunen (vt@james.hut.fi)
Jaakko Väyrynen (jjvayryn@james.hut.fi)

The target of the project is to create software which recognizes songs. The training set is a collection of MP3 songs by several artists. First feature extraction is done, then the GMM model for each song is taught. Also we will build models for each artist and genre. The models are time-independent which allows recognition from any part of the song. The system will be implemented in C.

Project #2: Survey of Weighted Finite State Transducers (WFST)

Teemu Hirsimäki (thirsima@james.hut.fi)

During the recent years, Weighted Finite State Transducers (WFST) have become an attractive framework for large vocabulary speech recognition. The increase in the computational power, and the development of efficient algorithms for composing and minimizing transducers have made it possible to build all information about phoneme models, context-dependency, pronunciation lexicon, and language model into transducers. With the WFST algorithms, these transducers can be composed, minimized and pruned efficiently off-line, before recognition, leading to a compact representation of the whole search network, which can be decoded with a simple Viterbi decoder.

The aim of this literature survey is to review papers dealing with the development of WFST algorithms that allowed to build real transducer-based recognition systems. In addition, the survey tries to answer the following two questions. (1) In what areas in the field of speech and language processing, transducers have been used beside speech recognition? (2) What are the current shortcomings of the WFST-based recognition framework, and what things must be done with other methods?

Project #3: Survey of Segment-Based Speech Recognition

Petri Korhonen (petri@acoustics.hut.fi)

The most modern speech recognition systems take as an input a set of features computed at fixed rate from short time windows of speech signal. An alternative framework to this is to prior to recognition phase to perform some acoustical analysis to get some explicit segmentation of the speech signal. The segments obtained this

way are of variable lengths, and different features can be used for different segments. The segmental framework allows a richer set of acoustic-phonetic features than can be incorporated into conventional frame-based representations. Systems based on this framework include for example SUMMIT. In this project I try to get an overview of this framework. Also I will try to find out what are the building blocks, and algorithms used in systems based on this framework.

Project #4: Speech Recognition System for XForms

Mikko Honkala (honkkis@tml.hut.fi)
Mikko Pohja (mpohja@tml.hut.fi)

XForms is a new abstract user interface definition language. In our opinion it could be used as the basis in multimodal services in the Web. In the project work, we will research and implement a speech recognition system for XForms. This system will allow filling forms and operating user interfaces written in XForms, using speech. We will use java-based Sphinx-4 speech recognition library (http://cmusphinx.sourceforge.net/sphinx4/). We will focus on how to implement the different form controls with the grammar-based recognition system. Also, we will implement the navigation within a form. The navigation differs from GUI based navigation, since in speech, the form has to be serialized. The implementation will be done to the XML browser X-Smiles (www.xsmiles.org). X-Smiles already has an XForms processor, and it is an open source project by the TML laboratory, so it is quite ideal environment to test multimodality and speech input. We will implement at least navigation, selection lists, and some data type-bound input fields in speech. Free form input fields are probably not feasible with grammar-based speech recognition system, such as Sphinx-4.

Project #5: Music recognition server based on GMMs

Pedro Díaz Jiménez (pdiaz@cc.hut.fi)

The aim of the project would be to develop a music recognition server. This server would function like existing CDDB servers, but will work with music files instead of Audio CDs. Users submit to the server some kind of parameters about the song (maybe observation vectors based on MFCCs?) and receive the title, author, etc. of the song. This server would be useful to set the metadata information (ID3 in mp3 files for example) of the user's music collection. If time permits I also would like to try to add author and genre identification features.

Project #6: An implementation of a token pass decoder

Janne Pylkkönen (jpylkkon@james.hut.fi)

The concept of token passing for speech recognition was introduced over 15 years ago. The most popular decoding technique for large vocabulary continuous speech

recognition (LVCSR) nowadays is the one-pass time-synchronous beam search strategy, which is still based on that same principle. The key advantage of token passing is the conceptually simple approach, which makes it easy to extend the strategy to handle many advanced problems in speech recognition, such as cross-word contexts and early language model pruning.

This work involves implementing a token pass decoder to the existing CIS-HUT LVCSR framework. It already contains a stack decoder, from which some of the code can be reused. However, the new decoder should take care of several problems previously unaddressed, such as the use of tied HMM states in the lexical prefix tree and the possibility to use cross-word triphone contexts. When finished, the efficiency of the token pass decoder approach will be compared to the existing stack decoder.

As an implementation of a decoder may prove to be quite extensive, a priority list of goals is defined. At the core is the construction of the lexical prefix tree and the implementation of the actual token pass algorithm for beam search. Next a problem of integrating early language model score based on bigrams is looked at. If time allows, also the implementation of cross-word triphone contexts is made. The overall design goal is to make the system easily extendable for future needs.

Project #7: Speech Recognition Based on Artificial Neural Networks

Veera Ala-Keturi (valaketu@cc.hut.fi)

Artificial neural networks (ANNs) are systems consisting of interconnected computational nodes imitating the human neurons. Neural networks can be used e.g. to approximate functions or classify data into similar clusters (in a supervised or unsupervised manner). I will first look at some basic theory of neural networks: perceptrons, multi-layer networks, feed-forward vs. recurrent networks, update criterions and algorithms etc. I will then study hybrid (connectionist) models where HMMs and NNs are used together in speech recognition. In a hybrid HMM/NN system the neural network estimates the posterior probabilities, which can greatly enhance the discrimination ability of the system. Finally I will look at the rather new field of research dealing with extracting features from the data by neural networks. The purpose of this survey is to obtain an understanding of the state-of-the-art in usage of neural networks for speech recognition, and find the pros and cons of each technique.

Project #8: Noise-Robust Speech Recognition

Bernhard Leiner (bernhard@footbag.at)

In my survey project I will try to give an overview about the different techniques and algorithms used to recognize speech in a noisy environment. The focus is on compensation of noise during the preprocessing stage (feature mapping) and model adaptation due to noise during the recognition.

Project #9: Language Modeling in Automatic Speech Recognition

Antti Puurula (Antti.Puurula@helsinki.fi)

Language models form an essential component in modern speech recognition system. Some of the missing percentages in ASR error rates could be related to inadequacy of the language models. The purpose of this survey project is to examine the modern n-gram models as well as some of the alternative approaches that have been tried.

Project #10: Experiments with Spoken Passphrase and Speaker Identification

Juha Raitio (juha.raitio@iki.fi)

The goal of this project work is to study the possibilities for spoken passphrase and speaker identification. Of interest is a system that would identify an utterance of a passphrase as a previously presented one by the same speaker, or discard it as unknown. The objective is to conduct background research on the challenges, in order to select a plausible approach, and conduct experiments by applying it. Optionally an on-line toy system is implemented. The effort should be documented in scientific manner.

Outline of experiments. K speakers utter M passphrases each N times, utterances are labeled by the speaker name and a passphrase id. T% of utterances of U% of passphrases by U% of the speakers are used for training a (speaker x passphrase x utterance) model. Utterances not used in training are used for testing. Each test passphrase must be either identified or discarded. Correct identifications and errors and their types are recorded and reported. Errors occur when an utterance is discarded though it should have been identified (type I), or if an utterance was identified incorrectly (type II). Type IIA error would occur is the speaker is identified incorrectly, IIB if the passphrase is identified incorrectly.

Research questions. Is it possible to control and balance the probabilities of type I and II errors by model selection? Optionally: what other factors e.g. passphrase length, number of passphrases M, number of speakers K, number of repetitions of the passphrase by a speaker N, etc. affect the success rate.

Project #11: Classifier Combination for Speech Recognition

Matti Aksela (matti.aksela@hut.fi)

For the actual course project, I will attempt to write up a survey of combination methods used in speech recognition. I will attempt to evaluate them with a more general view of classifier combining, and also consider the usability of some adaptive combination methods that have used in my previous research within the domain of speech recognition.

Project #12: Automatic Language Identification from Telephone Speech

Zhirong Yang (rozyang@cc.hut.fi)

The project work involves implementation and comparison of three approaches for automatic language identification of speech utterance: Gaussian mixture model (GMM) classification; single-language phone recognition followed by language-dependent, interpolated n-gram language modeling (PRLM); parallel PRLM, which uses multiple single-language phone recognizers, each trained in different language. Also, the performance by merging multiple language phone recognizers will be also investigated.