Special Lecture (406)
Spoken Language Dialog Systems
Speech Recognition

Rolf Schwitter
{schwitt}@ics.mq.edu.au
Today’s Program

• Speech recognition
• Phonemes
• Phases of ASR
• Acoustic models
• Grammars as language models
• N-grams as language models
• Recognition errors
• Excursus: SWI-Speech
Speech Recognition

- Speech produces a sound pressure wave which forms an acoustic signal.
- The microphone
  - receives the acoustic signal and
  - converts it to an analogue signal.
- To store the analogue signal, it must be converted to a digital signal.
- A speech recognizer tries to transform a digitally-encoded acoustic signal in a natural language into text in that language.
Automatic Speech Recogniser (ASR)

- Parameters used to characterize the capability of an ASR:

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaking Mode</td>
<td>Isolated words to continuous speech</td>
</tr>
<tr>
<td>Speaking Style</td>
<td>Read speech to spontaneous speech</td>
</tr>
<tr>
<td>Enrollment</td>
<td>Speaker-dependent to speaker-independent</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>Small (&lt;20 words) to large (&gt;20,000 words)</td>
</tr>
<tr>
<td>Language Model</td>
<td>Finite-state to context-sensitive</td>
</tr>
<tr>
<td>Perplexity</td>
<td>Small (&lt;10) to large (&gt;100)</td>
</tr>
<tr>
<td>SNR</td>
<td>High (&gt;30 dB) to low (&lt;10 dB)</td>
</tr>
<tr>
<td>Transducer</td>
<td>Voice-cancelling microphone to telephone</td>
</tr>
</tbody>
</table>
Speaking Mode

• Early speech recogniser engines required that the caller pause briefly between each word.
• These recognisers were said to have an isolated speaking mode.
• However, this mode of speaking is quite unnatural for people.
• Today, most speech recognisers use a continuous speaking mode.
• Continuous speaking mode is possible, since the recogniser automatically detects and segments words.
Speaking Style

- Some recognisers may be tuned to recognise speech that is read.
- For example: text spoken by a radio announcer.
- These recognisers are used to transcribe news programs.
- Spontaneous speech is uttered at the spur of the moment.
- Spontaneous speech may
  - not be grammatically correct
  - contain noise words such as "em" and "ah"
  - contain word fragments and repeated words.
Enrollment

- Enrollment is the process of training a speech recogniser.
- Some speech recognition engines require enrollment.
- A caller enrolls by speaking several words and phrases before using the system.
- The speech recogniser creates an acoustic model, that is a library of spoken sounds and their phonetic equivalents.
- Other speech recognisers come with pre-existing acoustic models and do not require training.
- Usually, enrollment improves the accuracy of the acoustic model.
Vocabulary

- Small vocabulary speech recognisers
  - only understand between twelve and a few hundred of words
  - are used for command and control applications
  - are quite fast and accurate.
- Large vocabulary speech recognisers
  - understand thousands of words
  - are used for dictation tasks
  - often fail in telephony environments (poor microphones).
Vocabulary

- Conversational speech recognisers
  - require the accuracy of a small vocabulary speech recogniser
  - but need the coverage of a large vocabulary speech recogniser
  - are used for voice portals
  - have different subsets of the vocabulary active at different points during the application.
Language Model

- A language model is a representation of
  - words from a specific language
  - and their corresponding phonemes.
- The simplest language models are finite-state networks (where permissible words following each word are given explicitly).
- More general language models are specified in terms of context-sensitive grammar.
Perplexity

• Perplexity is a measure for the average number of words a caller can speak after the current word.

• For example, after speaking the words
  – *turn the volume* . . .
  the caller may only speak two words
  – *up* or *down*.

• Here the perplexity is 2.
Perplexity

• If the caller can speak the names of 300 travel destinations after
  – I want to travel to …
  then the perplexity is 300.

• In general, accuracy declines as perplexity increases.
**Word Error Rate**

- The accuracy of a speech recognition engine is measured by its word error rate:
  - the number of spoken words missing in the text
  - plus the number of words in the text that were not spoken
  - plus the number of words in the text that do not match the corresponding spoken words
  - divided by the number of words the caller spoke.

- Generally, speaker-independent systems have a word error rate about four times that of a speaker-dependent system.
Problems of Describing Speech

- Sentences consist of sequences of words, in text these are delimited by spaces.
- When we produce speech, there are no markers for word boundaries.
- Speech can be described as a sequence of phonemes.
- To identify words, we need to search for an interpretation within the phoneme sequence.
Examples

• Word boundary
  – It’s easy to recognize speech.
  – It’s easy to wreck a nice beach.
  – I scream for ice-cream.
  – Ice-cream for ice-cream.

• Homophones
  – The results of the first period are explained in the next paragraph.
  – period is not a punctuation mark
  – paragraph is not a formatting command.
Phonemes

• A phoneme is the smallest segment of sound that can distinguish two words.
• Two words that are differentiated by one phoneme, such as /cat/ → /rat/
  are known as a minimal pair.
• English has around 45 phonemes.
Phonemes

- Phonemes are classified into the following broad categories:
  - Vowels
    - monophthongs (hat)
    - diphthongs (height)
  - Consonants
    - stops (pat)
    - nasales (sing)
    - fricatives (thing)
Phonetics

- Phonemes are abstract units: a useful way of describing speech.
- They provide a categorical description of speech sounds.
- In real speech, there are no hard boundaries between the categories: we use context to disambiguate.
  - *He was a good cook.*
  - *We knew he could cook.*
- Each phoneme can be realised in different ways depending on context and speaker → allophones.
Variability in Speech

- The main problem for speech recognition is coping with the variability of the speech signal.
- Phonemes vary with context, speaking rate, accent, ...
- Words have alternate pronunciations and pronunciations can change based on speaking rate:
  Photograph: /fotograf/ or /fod@graf/
- Individual speakers vary in their physiology, accent, state of health.
- All of this combines to mean that no two utterances of a word or phrase will look exactly alike.
Phases of ASR

Audio Input

Feature Extraction

Phoneme Identification

Acoustic Model

Word Identification

Language Model

Words & Phrases
Feature Extraction

- Spoken speech is converted to frames of a few milliseconds each.
- Frames are then converted to digital form via feature extraction.
- Each frame represented as a feature vector, containing frequency and energy information: e.g. one vector of 30 numbers every 100 ms.
- The goal of feature extraction is to:
  - remove the parts of the signal that don’t contribute to phonetic identity (e.g. pitch, transmission line noise)
  - reduce the amount of data that the pattern matcher has to process.
Phoneme Identification

- An **acoustic model** is a mapping from speech features to phonemes.
- These mappings are modified when a caller trains a recogniser.
- **Phoneme identification** is the process that uses the acoustic model to transform extracted features uttered by the caller to sequences of phonemes.
- Two technologies are used to achieve this mapping
  - Hidden Markov Models
  - Neural Networks
Acoustic Models

- A different acoustic model will be required for each different dialect.
- Acoustic models available at Nuance web site:
  
  North American English, UK English, Australian/NZ English, Singapore English, South African English, French, German, Portuguese for Brazil, Swedish, Japanese, Mandarin/Taiwan, Swiss German, Cantonese, Canadian French, Norwegian, Spanish American, Italian, Czech, ...
Hidden Markov Model (HMM)

• A HMM is a statistical pattern matcher:
  – each state represents a phoneme
  – each transition represents the probability of two phonemes occurring in sequence.

• When features are fed into a HMM,
  – vectors of probabilities representing the likelihood
  – that the caller spoke each of the possible phonemes
  – emerge from the model.

• HMM are most widely used in commercial ASR.
Neural Networks

- Each node of a neural network corresponds to one of the phonemes.
- Each node
  - examines speech features extracted from the caller’s utterance
  - determines how close they correspond to a particular phoneme
  - assigns a score.
- The scores of all neural network nodes are combined.
- The output is a vector of probabilities representing the likelihood that the caller spoke each of the possible phonemes.
Word Identification

- **Word identification** is the process of mapping sequences of phonemes (vectors representing probabilities of phoneme sets) to actual words.
- This process consists of searching a language model.
- Sequences of phonemes are mapped to words in the language model.
- What are the words in the language model?
- Usually, a subset of words of a language used in a specific domain.
- Two kind of language models are common:
  - grammar-based models: specify alternatives rules
  - statistical models: trained on a large body of text.
A Fragment of a Suburbs Dictionary

- abbotsbury: ab*tsb*ri
- abbotsbury: ab*tsb*ri
- abbotsford: ab*tsf*d
- abbotsford: ab*tsfOd
- acacia_gardens: *keSi*gAd*nz
- acacia_gardens: *keSi*gAd*nz
- acacia_gardens: *keS*gAd*nz
- acacia_gardens: *keS*gAd*nz
- acacia_gardens: *keSi*gAdnz
- acacia_gardens: *keSi*gAdnz
Grammars as Language Models

- Small vocabularies of words are usually described by grammars.
- A grammar is one kind of language model.
- A grammar defines the possible sequences of words that can be recognised at some point in the dialog.
- A grammar compiler
  - expands the grammars into a list of acceptable words and phrases
  - extracts the sequence of phonemes for each of these words.
- The words and sequences of phonemes are placed in a language model data structure that can be searched quickly by the ASR.
Ways of Helping the Recogniser

• Specify a grammar for each dialog state.
• Make some part of a grammar more likely than others by adding probabilities, for example:
  – [fish~10 (angel fish)~2 anthia~0.1]
• Dynamically refine the grammar: e.g. populate airport city grammars on the basis of the country under discussion.
How Many Words?

- A typical desk dictionary contains around 50000-150000 entries.
- In 44 million words of Associated press newswire text collected over 10 months, there were 300000 different word types.
- It has been suggested that by the age 17 we know 80000 words.
- Your active vocabulary may be ~20000 words while your passive vocabulary is ~100000 words.
- It has been estimated that foreign students need more than 20000 words as passive vocabulary and at least 10000 as active vocabulary to attend college in an American university.
Observation

- Not all words are equally likely.
- It has been estimated that 8000 morphemes is sufficient to handle 95% of texts.
- Typically, 15 most frequent words account for 25% of tokens.
- 100 most frequent words account for 60% of tokens.
- So, augment recognition search space with probabilities.
## Word Frequencies

<table>
<thead>
<tr>
<th>Rank</th>
<th>Spoken English</th>
<th>Written English</th>
<th>French</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>the</td>
<td>de</td>
<td>der</td>
</tr>
<tr>
<td>2</td>
<td>and</td>
<td>of</td>
<td>le</td>
<td>die</td>
</tr>
<tr>
<td>3</td>
<td>I</td>
<td>to</td>
<td>la</td>
<td>und</td>
</tr>
<tr>
<td>4</td>
<td>to</td>
<td>in</td>
<td>et</td>
<td>in</td>
</tr>
<tr>
<td>5</td>
<td>of</td>
<td>and</td>
<td>les</td>
<td>des</td>
</tr>
<tr>
<td>6</td>
<td>a</td>
<td>a</td>
<td>des</td>
<td>den</td>
</tr>
<tr>
<td>7</td>
<td>you</td>
<td>for</td>
<td>est</td>
<td>zu</td>
</tr>
<tr>
<td>8</td>
<td>that</td>
<td>was</td>
<td>un</td>
<td>das</td>
</tr>
<tr>
<td>9</td>
<td>in</td>
<td>is</td>
<td>une</td>
<td>von</td>
</tr>
<tr>
<td>10</td>
<td>it</td>
<td>that</td>
<td>du</td>
<td>für</td>
</tr>
</tbody>
</table>
N-Grams as Language Model

- Basic idea:
  - the probability of occurrence of a symbol is conditioned upon the prior occurrence of $N-1$ other symbols.
  - typically constructed from statistics obtained from a large corpus of text using the co-occurrences of words in the corpus to determine word sequence probabilities.

- $N$ is typically 2 or 3.
The Predictive Power of N-Grams

- Example input:
  - Yesterday I went to the ...

- Bigrams:
  - Next word is something that can or is likely to follow ‘the’

- Trigrams:
  - Next word is something that can or is likely to follow ‘to the’
Word N-Gram Models

- **Bigram model:**
  - Gives the probability of all possible next words on the basis of the current word only
  - \( p(w_2 | w_1) \)

- **Trigram model:**
  - Based on two words of context
  - \( p(w_3 | w_1w_2) \)
Why use N-Gram Language Modelling?

- Allow for large vocabulary applications.
- A context free grammar of reasonable complexity can never foresee all the different utterance patterns that callers may use in spontaneous speech input.
Class N-Gram Models

• Group words into classes before computing statistics.
• Possible classes
  – syntactic categories such as nouns or verbs
  – semantic categories such as people’s name, city names, days of the week …
  – automatically-learned classes.
Class N-Gram Models

• Advantage
  – less data to compute
  – accommodates sparse data and limitations of corpus size

• Disadvantage
  – may lose important distinctions
Grammar Based Models versus N-Grams

- **Grammars**
  - typically constructed by hand
  - difficult to maintain as they get larger
  - may be overly restrictive
- **N-gram models**
  - computed automatically
  - more scope for misrecognition
  - how do you extract the semantics?
Language Model (LM) Creation

Diagram:

Grammar → Compiler → LM

Lexicon

Text → Statistical Routines → N-gram → Compiler → LM

Lexicon
Recognition Errors

• Speech recognition errors will always occur.
• Therefore:
  – Tune prompts to lead the caller to say words that are covered by the grammar.
  – Write error handlers to help callers recover from <noinput> and <nomatch> errors.
  – Test the application before deployment, analyse and resolve frequent errors.
Errors and Fixes

- The caller speaks one or more words that are not in the grammar.
  → Revise the prompt or extend the grammar.
- The words in the grammar sound alike.
  → Change similar sounding words in the grammar.
- The caller speaks with disfluency.
  → Design prompts so that the caller responds with individual words.
Errors and Fixes

• The caller has an accent.
  → Use an acoustic model designed for the accent.
  → Regress to DTMF dialogs for callers with strong accent.
• The caller has a cold, is intoxicated, or is extremely tired.
  → Encourage the caller to try another day.
• The caller makes non-speech sounds (laughing or coughing).
  → Encourage the caller to mute the phone when laughing.
Errors and Fixes

- The caller speaks with another person.
  → Encourage the caller to mute the phone.
- The caller is not able to complete an utterance.
  → Adjust the time-out parameter and allow for longer pauses.
Excursus: SWI-Speech

- SWI-Speech is an interface between SWI-Prolog and Microsoft SAPI.
- Available at http://www.ai.uga.edu/~mc/ProNTo/
- The SWI-Speech package consists of two files: SWI_Speech.dll and SWI_Speech.pl.
- Load SWI_Speech.pl into Prolog as follows:
  
  ```prolog
  :- ensure_loaded('swi_speech.pl').
  ?- listen(Sentence).
  Sentence = 'Hello World'
  ?- speak('Hello World').
  ```
Excursus: SWI-Speech

• Attributes for speech output:
  - `set_rate(0).` % number between -10 and 10
  - `set_volume(100).` % number between 0 and 100
  - `speak_with_attributes('Hello world').`

• Age and gender attribute:
  - `set_age(adult,required).` % child, teen, adult or senior
  - `set_gender(male,optional).` % male or female
  - `speak_with_attributes('Hello world').`
Take-home Messages

• Speech contains variability from many sources, most of which make ASR more difficult.

• Feature extraction is the process to convert the acoustic signal to a (simplified) digital format.

• Phoneme identification is the process that transforms the extracted features via an acoustic model to sequences of phonemes.

• Word identification is the process that maps sequences of phonemes via a language model to actual words.

• Language models can be grammar-based or n-gram based.