Lexicon and Pronunciations

Steve Renals

Automatic Speech Recognition – ASR Lecture 11
2 March 2017

Mathematical framework

HMM Framework for speech recognition. Let W be any possible transcription, and X be the observed acoustics; then we want to find the most probable transcription W^*:

$$
\begin{aligned}
W^* &= \arg\max_W P(W \mid X) \\
&= \arg\max_W \frac{P(X \mid W) P(W)}{P(X)} \\
&= \arg\max_W P(X \mid W) P(W)
\end{aligned}
$$

Mathematical framework

HMM Framework for speech recognition. Let W be any possible transcription, and X be the observed acoustics; then we want to find the most probable transcription W^*:

$$\begin{aligned} W^* &= \arg\max_W P(W \mid X) \\ &= \arg\max_W \frac{P(X \mid W)P(W)}{P(X)} \\ &= \arg\max_W P(X \mid W)P(W) \end{aligned}$$

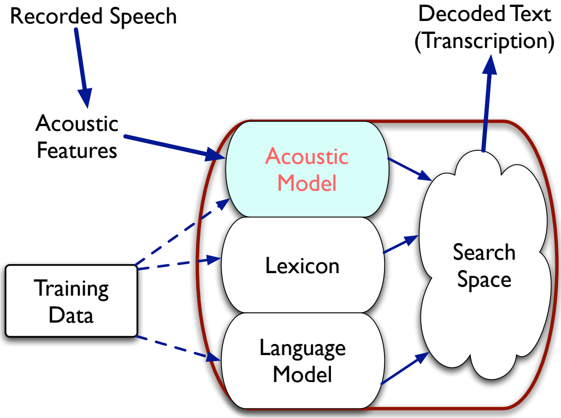
Words are composed of a sequence of HMM states Q:

$$\begin{aligned} W^* &= \arg\max_W P(X \mid Q, W)P(Q, W) \\ &\simeq \arg\max_W \sum_Q P(X \mid Q)P(Q \mid W)P(W) \\ &\simeq \arg\max_W \max_Q P(X \mid Q)P(Q \mid W)P(W) \end{aligned}$$

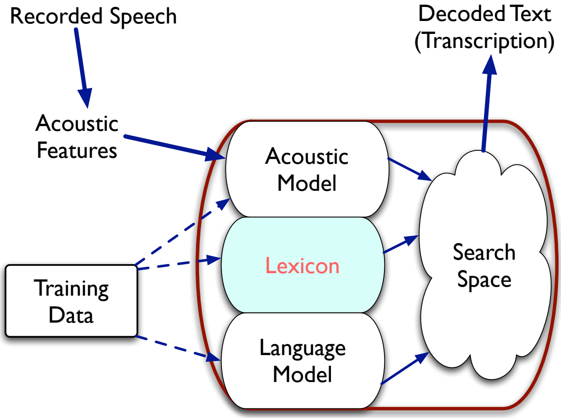
Three levels of model

- **Acoustic model** $P(X \mid Q)$
  Probability of the acoustics given the phone states: context-dependent HMMs using state clustering, phonetic decision trees, etc.
- **Pronunciation model** $P(Q \mid W)$
  Probability of the phone states given the words; may be as simple a dictionary of pronunciations, or a more complex model
- **Language model** $P(W)$
  Probability of a sequence of words. Typically an n-gram

HMM Speech Recognition



HMM Speech Recognition



Pronunciation dictionary

- Words and their pronunciations provide the link between sub-word HMMs and language models
- Written by human experts
- Typically based on phones

Pronunciation dictionary

- Words and their pronunciations provide the link between sub-word HMMs and language models
- Written by human experts
- Typically based on phones
- Constructing a dictionary involves
  1. Selection of the words in the dictionary—want to ensure high coverage of words in test data
  2. Representation of the pronunciation(s) of each word

Pronunciation dictionary

- Words and their pronunciations provide the link between sub-word HMMs and language models
- Written by human experts
- Typically based on phones
- Constructing a dictionary involves
  1. Selection of the words in the dictionary—want to ensure high coverage of words in test data
  2. Representation of the pronunciation(s) of each word
- Explicit modelling of pronunciation variation

Out-of-vocabulary (OOV) rate

- OOV rate: percent of word tokens in test data that are not contained in the ASR system dictionary
- Training vocabulary requires pronunciations for *all* words in training data (since training requires an HMM to be constructed for each training utterance)
- Select the recognition vocabulary to minimize the OOV rate (by testing on development data)
- Recognition vocabulary may be different to training vocabulary
- Empirical result: each OOV word results in 1.5–2 extra errors ($>$1 due to the loss of contextual information)

Multilingual aspects

- Many languages are morphologically richer than English: this has a major effect of vocabulary construction and language modelling
- Compounding (eg German): decompose compund words into constituent parts, and carry out pronunciation and language modelling on the decomposed parts
- Highly inflected languages (eg Arabic, Slavic languages): specific components for modelling inflection (eg factored language models)
- Inflecting and compounding languages (eg Finnish)
- All approaches aim to reduce ASR errors by reducing the OOV rate through modelling at the morph level; also addresses data sparsity

Single and multiple pronunciations

- Words may have multiple pronunciations:
  1. Accent, dialect: *tomato*, *zebra*
     global changes to dictionary based on consistent pronunciation variations
  2. Phonological phenomena: *handbag*/ `h ae m b ae g`
     I can't stay / `[ah k ae n s t ay]`
  3. Part of speech: *project*, *excuse*

Single and multiple pronunciations

- Words may have multiple pronunciations:
  1. Accent, dialect: *tomato*, *zebra*
     global changes to dictionary based on consistent pronunciation variations
  2. Phonological phenomena: *handbag*/ `h ae m b ae g`
     I can't stay / `[ah k ae n s t ay]`
  3. Part of speech: *project*, *excuse*
- This seems to imply many pronunciations per word, including:
  1. Global transform based on speaker characteristics
  2. Context-dependent pronunciation models, encoding of phonological phenomena

Single and multiple pronunciations

- Words may have multiple pronunciations:
  1. Accent, dialect: *tomato*, *zebra*
     global changes to dictionary based on consistent pronunciation variations
  2. Phonological phenomena: *handbag*/ `h ae m b ae g`
     I can't stay / `[ah k ae n s t ay]`
  3. Part of speech: *project*, *excuse*
- This seems to imply many pronunciations per word, including:
  1. Global transform based on speaker characteristics
  2. Context-dependent pronunciation models, encoding of phonological phenomena
- **BUT** state-of-the-art large vocabulary systems average about 1.1 pronunciations per word: most words have a single pronunciation

Consistency vs Fidelity

- Empirical finding: adding pronunciation variants can result in reduced accuracy
- Adding pronunciations gives more “flexibility” to word models and increases the number of potential ambiguities—more possible state sequences to match the observed acoustics

Consistency vs Fidelity

- Empirical finding: adding pronunciation variants can result in reduced accuracy
- Adding pronunciations gives more "flexibility" to word models and increases the number of potential ambiguities—more possible state sequences to match the observed acoustics
- Speech recognition uses a consistent rather than a faithful representation of pronunciations
- A consistent representation requires only that the same word has the same phonemic representation (possibly with alternates): the training data need only be transcribed at the word level
- A faithful phonemic representation requires a detailed phonetic transcription of the training speech (much too expensive for large training data sets)

Modelling pronunciation variability

- State-of-the-art systems absorb variations in pronunciation in the acoustic models
- Context-dependent acoustic models may be though of as giving broad class representation of word context
- Cross-word context dependent models can implicitly represent cross-word phonological phenomena
- Hain (2002): a carefully constructed single pronunciation dictionary (using most common alignments) can result in a more accurate system than a multiple pronunciation dictionary

Current topics in pronunciation modelling

- Automatic learning of pronunciation variations or alternative pronunciations for some words – e.g. learning probability distribution over possible pronunciations generated by grapheme-to-phoneme models
  - Automatic learning of pronunciations of new words based on an initial seed lexicon
- Joint learning of the inventory of subword units and the pronunciation lexicon
- Sub-phonetic / articulatory feature model
- Grapheme-based modelling: model at the character level and remove the problem of pronunciation modelling entirely