Speech Recognition for Under-resourced Languages using Probabilistic Transcriptions

Preethi Jyothi
Department of CSE, IIT Bombay

CS344 Guest Lecture
February 7, 2017
Automatic speech recognition (ASR):
Translate spoken words into text
ASR isn’t to blame for this...

“Siri I'm bleeding really bad can you call me an ambulance”

From now on, I’ll call you ‘An Ambulance’. OK?
Automatic speech recognition (ASR): Translate spoken words into text

Modern ASR systems are dominated by statistical methods pioneered by [Jelenik ’76]
Standard ASR Pipeline

$p(X|Q) \equiv \frac{p(X|Q,W)}{p(W)}$

ACOUSTIC MODEL

$p(Q|W)$

PRONUNCIATION MODEL

$p(W)$

LANGUAGE MODEL

$p(X,Q,W)$

Decoding: Given $X$, find $\arg\max_W p(W|X)$

$= \arg\max_W p(W,X)$

$= \arg\max_W \sum_W p(X,Q,W)$
ASR over the years

- Great progress in ASR performance
- Aided by algorithmic and computational advances
- Recently: Baidu’s Deep Speech 2 comparable to human performance
- Trained on about 10,000 hours of labeled speech
- But limited language diversity

Languages with ASR

E.g., Google Voice Search supports < 80 out of 7000 languages

ASR for all languages

- ASR systems in all languages?
  - Speech is the primary means of human communication
  - Develop natural interfaces for both literate & illiterate users
  - Contribute to preservation of endangered languages
Lack of Transcribed Corpora

• Major challenge: Building ASR systems is very data-hungry
  • Require large amounts of labeled speech data: Speech audio with *matching transcriptions*
  • Transcription by native speakers is a laborious and expensive process

• Crowdsourcing might help alleviate the problem
  • However, *significant* mismatch in native languages of crowd workers and native language populations in the world
Native Language Mismatch

- Very few (to zero) crowd workers speak minority languages
- Distributional mismatch between language background of crowd workers with the language expertise required to complete transcription tasks
Mismatched Crowdsourcing

- A major bottleneck for ASR in new languages: Labeled speech
- Transcribers need to be native speakers

Use Non-native Speakers?

Mismatched Crowdsourcing

How can it possibly work?!\(^1\)

\(^1\)[Jyothi & Hasegawa-Johnson AAAI-15 & Interspeech-15]
Mismatched Crowdsourcing

- How can it possibly work?! [Best ’94, Flege ’95, etc.]
  - We are typically bad at perceiving speech in foreign languages!
    - Unfamiliar sounds, no vocabulary, no language model to go by, distorted by native languages, ...
Mismatched Crowdsourcing

• How can it possibly work?! [Best ’94, Flege ’95, etc.]

  • We are typically bad at perceiving speech in foreign languages!

    • Unfamiliar sounds, no vocabulary, no language model to go by, distorted by native languages, ...

• An extremely noisy channel
Mismatched Crowdsourcing

- An extremely noisy channel
Solution: Error Correction

- Learn channel characteristics of the foreign listener
Solution: Error Correction

- Learn channel characteristics of the foreign listener

A Probabilistic Finite State model

Trained using the Expectation-Maximization (EM) algorithm

\[ x = [tʃ] [iː] [k] \]

\[ p \left( \text{“chieg”} \mid x \right) = 0.11 \]

\[ p \left( \text{“cheek”} \mid x \right) = 0.25 \]
Conditional probability of English letters given Hindi phones
Solution: Error Correction

- Learn channel characteristics of the foreign listener
- But also need to use an error-correcting code

But how can we encode speech?
Solution: Error Correction

- Learn channel characteristics of the foreign listener
- Use a repetition code!

![Diagram showing the encoding and decoding process with an example message in Hindi and English.](image-url)
Solution: Error Correction

- Learn channel characteristics of the foreign listener
- Use a repetition code!
Example

pūrā

puda
waatap
fotup
puddop
pooda
puda
purap
poduck
purap
foodap

paḍegā
paḍegā
pratī
paḍā
pūrā
pūrā
pūrā
pūrā
pūrā
Labeling Error Rates

- Impressive accuracy (~5% error) on a medium-vocabulary isolated-word task!

![Graph showing error rates for 1-best and 2-best repetitions.](Jyothi & Hasegawa-Johnson AAAI-15)
Information-theoretic Analysis

- Conditional entropy, \( H(X | Y) \), of the spoken words (Hindi), \( X \), given the crowd transcripts \( Y \), captures the amount of information lost in transmission.

- \( H(X | Y) \) can be naively upper-bounded using corpus cross-entropy.

- However, errors in our channel model accumulate with increase in the number of repetitions, resulting in this upper-bound becoming less tight.
Information-theoretic Analysis

Tighter bound on $H(X \mid Y)$ using an auxiliary random variable, $Z \in \{0,1\}$

Consider $W = \epsilon$ when $Z = 0$, $W = X$ when $Z = 1$
We set $Z = 1$ when $q(x\mid y)$ is sufficiently low

$W$ represents side channel information indicating when the model needs to be corrected

$$H(X \mid Y) \leq p_0 \cdot H(X \mid Y, Z=0) + H(Z) + (1 - p_0) \log |X|$$

where $p_0 = p(Z=0)$ and $X$ is the input alphabet

Upper-bounded using corpus cross-entropy
Our upper-bound estimates for CE are clearly tighter than the naive cross-entropy upper-bound estimate.
Mismatched Crowdsourcing for Continuous Speech
Continuous Speech

Decode each transcript individually using Maximum Likelihood Decoding and pick the output with the best score

Exact Maximum-Likelihood Decoding of multiple strings is intractable for long utterances

Word error rate: 77%
Continuous Speech

Can we do better?
An outlier with a good score shouldn’t be chosen over what many similar looking transcripts predict

Decode each transcript individually using Maximum Likelihood Decoding and pick the output with the best score

Word error rate: 77%
Continuous Speech

Data Filtering:
Use an edit-distance based similarity metric to discover a “cluster” to retain.

Can we do better?
An outlier with a good score shouldn’t be chosen over what many similar looking transcripts predict.
<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>v</td>
<td>i</td>
<td>e</td>
<td>m</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>b</td>
<td>i</td>
<td>a</td>
<td>n</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>v</td>
<td>e</td>
<td>a</td>
<td>m</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>v</td>
<td>e</td>
<td>a</td>
<td>n</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>b</td>
<td>i</td>
<td>a</td>
<td>n</td>
</tr>
</tbody>
</table>

Diagram: Nodes labeled 1 to 5 interconnected.
Continuous Speech

Data Filtering:
Use an edit-distance based similarity metric to discover a “cluster” to retain.

Can we do better?
An outlier with a good score shouldn’t be chosen over what many similar looking transcripts predict

Word error rate: 68%
Continuous Speech

Data Filtering: Use an edit-distance based similarity metric to discover a “cluster” to retain.

Can we do even better?

Word error rate: 68%
Channel Merger

Data Filtering
Discover “typical” transcripts

Alignment
NP-hard!
Approximation via incremental alignment algorithm

Independent Invocations of CHANNEL

ENCODER
Speaker
Repeat
Merge
Merged Channel Decoder

ENCODER

Data Filtering

Alignment
keeajaga
giyajayga
keeajaygah
chajijega

kiyā jāyegā
<table>
<thead>
<tr>
<th>k</th>
<th>E</th>
<th>-</th>
<th>a</th>
<th>j</th>
<th>a</th>
<th>g</th>
<th>a</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>g</td>
<td>i</td>
<td>y</td>
<td>a</td>
<td>j</td>
<td>Y</td>
<td>g</td>
<td>a</td>
<td>-</td>
</tr>
<tr>
<td>k</td>
<td>E</td>
<td>-</td>
<td>a</td>
<td>j</td>
<td>Y</td>
<td>g</td>
<td>a</td>
<td>h</td>
</tr>
<tr>
<td>C</td>
<td>-</td>
<td>-</td>
<td>Y</td>
<td>j</td>
<td>e</td>
<td>g</td>
<td>a</td>
<td>-</td>
</tr>
</tbody>
</table>

**kiyā jāyegā**
Channel Merger

Independent Invocations of CHANNEL

Data Filtering
Discover “typical” transcripts

Alignment
NP-hard!
Approximation via incremental alignment algorithm

Merge
Merge into one probabilistic transcript
### Merge Transcripts

<table>
<thead>
<tr>
<th>k</th>
<th>E</th>
<th>-</th>
<th>a</th>
<th>j</th>
<th>a</th>
<th>g</th>
<th>a</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>g</td>
<td>i</td>
<td>y</td>
<td>a</td>
<td>j</td>
<td>Y</td>
<td>g</td>
<td>a</td>
<td>-</td>
</tr>
<tr>
<td>k</td>
<td>E</td>
<td>-</td>
<td>a</td>
<td>j</td>
<td>Y</td>
<td>g</td>
<td>a</td>
<td>h</td>
</tr>
<tr>
<td>C</td>
<td>-</td>
<td>-</td>
<td>Y</td>
<td>j</td>
<td>e</td>
<td>g</td>
<td>a</td>
<td>-</td>
</tr>
</tbody>
</table>

*kiyā jāyegā*
Channel Merger

Independent Invocations of CHANNEL

Data Filtering
Discover “typical” transcripts

Alignment
NP-hard!
Approximation via incremental alignment algorithm

Merge
Merge into one probabilistic transcript
Model for merged channel
Channel Merger

Data Filtering
Discover “typical” transcripts

Alignment
NP-hard! Approximation via incremental alignment algorithm

Merge
Merge into one probabilistic transcript

Model for merged channel

Shortlist & Decode
List Decoding + Exact Decoding from List
Probabilistic Transcriptions

Tacapo piza
strucka po zapecham
trakapo trabiza
Straka pose ta peesome
straka po ta pisha
strah kah poh chah peesh um
chaka-pu shapisha
stakkappoo sabeesham
takapo chapiser
Strike a pose some pizza

\[1\text{[P. Jyothi & Hasegawa-Johnson, }\textit{Interspeech-15}\]
Probabilistic Transcriptions

<table>
<thead>
<tr>
<th>stra</th>
<th>kapo</th>
<th>ztata</th>
<th>pe</th>
<th>som</th>
</tr>
</thead>
<tbody>
<tr>
<td>stra</td>
<td>kapo</td>
<td>tata</td>
<td>pi</td>
<td>sa</td>
</tr>
<tr>
<td>strahkapo</td>
<td>Cahp</td>
<td>E</td>
<td>Sum</td>
<td></td>
</tr>
<tr>
<td>strY</td>
<td>kapozs</td>
<td>amp</td>
<td>Etsa</td>
<td></td>
</tr>
</tbody>
</table>

Probabilistic phone-based transcriptions derived from alignments¹

¹[P. Jyothi & Hasegawa-Johnson, Interspeech-15]
Transcription Error Rates

Error Rates (%)

- No data filtering
- With data filtering
- Phone Error Rates

44% drop!

[Jyothi & Hasegawa-Johnson *Interspeech*-15]
Adapting ASR Systems using Mismatched Transcriptions

February 7, 2017
Next Step?

- Respectable accuracy from mismatched transcriptions
  - But can this be leveraged for building ASR systems?

- Plan: Baseline ASR trained on *other* languages will be *adapted* using mismatched transcriptions
  - Baseline could use data-hungry technology like Deep Neural Networks (DNNs)

- Project at 2015 Jelenik Summer Workshop [JSALT ’15]
  - Several languages considered: Hungarian, Mandarin, Swahili etc.
More than meets the eye

- Mismatched transcripts too noisy to be used in the traditional way for ASR training
  - Use as probabilistic transcripts
  - Measuring additional information in probabilistic transcripts:
    - How error rates fall when more “advice” is made available to the decoder

![Graph showing phone error rates vs. advice per phone](image)

- Hungarian
- Mandarin
- Swahili
ASR Systems for Comparison

- **Multilingual**: Train on 6 languages (Arabic, Cantonese, Dutch, Hungarian, Mandarin, Urdu) and test on a new target language (Swahili).

- **Semi-supervised DNN**: Transcribe unlabeled audio from the target language using a DNN-based multilingual ASR system and use it to further re-train the DNN models.
Mismatched Transcriptions for ASR

Phone Error Rates (%)

- Multilingual
- Semi-supervised DNN
- + Adaptation

Swahili
Mandarin
Hungarian

25% drop!

[Liu*, Jyothi* et al. ICASSP’ 16]
Native Language Backgrounds of Mismatched Crowds

February 7, 2017
Channel Selection

• Can we do better by selecting the language background of the transcribers?

• How should we select the transcribers?

  • Understanding when phones get misperceived

  • How is this correlated with transcribers’ language background
Understanding the Mismatched Channel

- When are two phones confused with each other?
  - If they are "phonologically close" to each other

- Phonological distance between two phones
  - Use *distinctive features* (DF) from linguistic theory [Chomsky & Halle, ’68] [Phoible ’15]
  - Phones as "code words" in the DF-space. Hamming distance measures phone contrast.

37 DFs:
- nasal
- tone
- sonorant
- labial
- trill
- front
- back
Distance Distribution of the Code

[Varshney, Jyothi & Hasegawa-Johnson /TA-16]
Codes for different languages exhibit similar distributions.
**Understanding the Mismatched Channel**

- **Hypothesis**: Phonological distance in the DF-space correlated with *phone confusion* in the mismatched channel.

- Phone confusion quantified using the **total variational distance** between the output distributions of the *channel*.

- We call it **phone-pair distinction**.
Phone Pair Distinction vs. DF-Distance

![Graph showing the relationship between phone pair distinction by English speakers and DF-distance, with color coding for frequency and a trend line.](image)
Phone Pair Distinction

- Phone pair distinction by Mandarin speakers
- Phone pair distinction by English speakers

Frequency (log scale)
Phone Pair Distinction

- Clearly, DF-distance positively correlated with phone-pair distinction
- Difference across native language backgrounds
  - Different DFs are prominent in different languages
Phone Pair Distinction

- Clearly, DF-distance positively correlated with phone-pair distinction
- Difference across native language backgrounds
  - Different DFs are prominent in different languages
  - Ongoing work: A model that takes into account DF presence/prominence
Summary

- ASR for low-resource languages presents challenging research problems

- In this talk:
  - Establish the possibility of acquiring speech transcriptions using mismatched crowds
  - Demonstrate the impact of mismatched transcriptions on ASR performance
  - Investigate relation of transcriber native languages with phone confusion

- Future research: Optimally select mismatched transcribers to further improve impact of mismatched transcriptions
Summary

ASR for low-resource languages presents challenging research problems

In this talk:

- Establish the possibility of acquiring speech transcriptions using mismatched crowds
- Demonstrate the impact of mismatched transcriptions on ASR performance
- Investigate relation of transcriber native languages with phone confusion

1Based on joint works with Mark Hasegawa-Johnson, Lav Varshney and participants at the 2015 Jelinek Summer Workshop.