

# Voice Search

*Opportunities, Challenges, and Methods*



8 October 2014

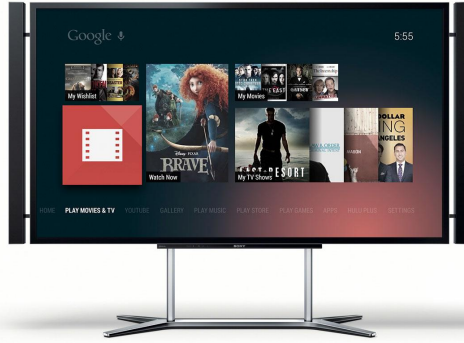
PROPOR 2014 - Sao Carlo, Brazil



# Opportunities

Speech is core to the future  
of mobile

---

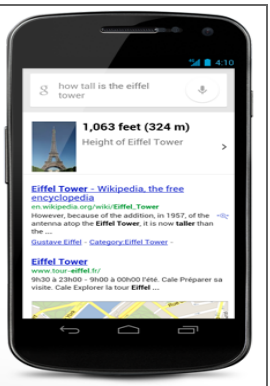
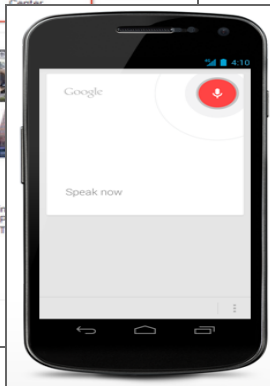
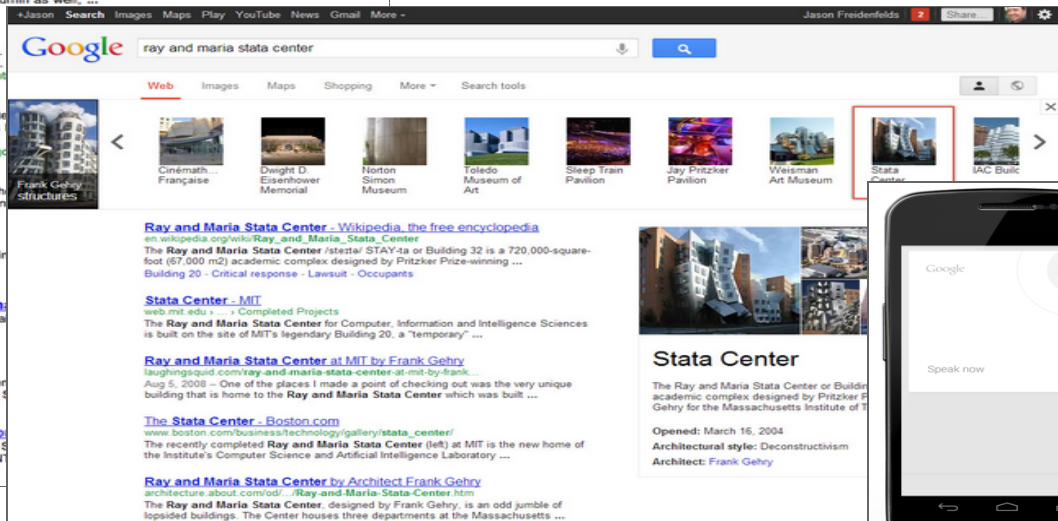
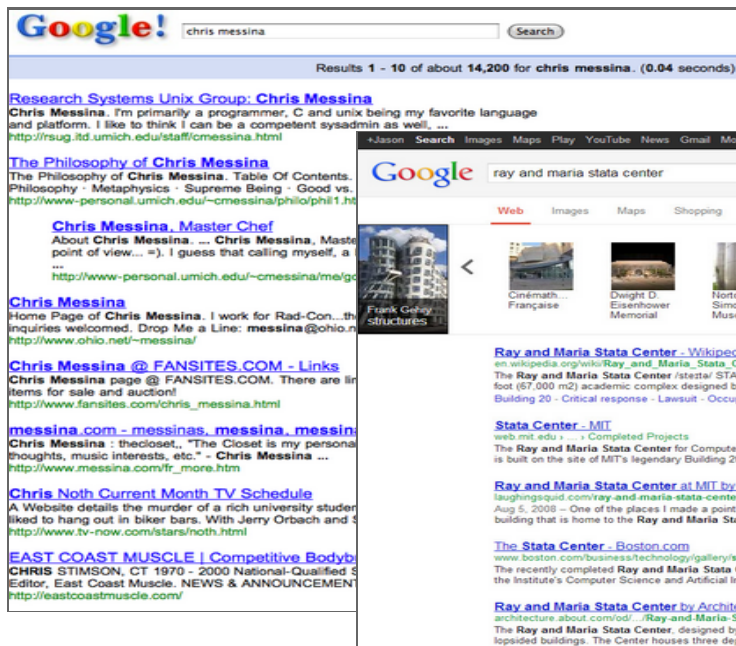


Google





# Evolution of Search



10 blue links



more than just keywords



mobile, voice

# Google Voice search with spoken answers

what will the weather be like in Monterey on Saturday

Monterey, CA

12°

CLEAR  
3m/s  
10%

TUE	WED	THU	FRI
13° 7°	13° 9°	14° 9°	16° 10°

Google web results

Monterev Weather Forecast and

Web Images Places

what movies are playing this weekend

**The Host**

Scifi, Fantasy, ... | PG-13 | 2hr 5min

Play trailer

What if everything you love was taken from you in the blink of an eye? "The... [more >](#)

★★★★★ (5)

Century Cinemas 16	0.3 mi
11:00am 1:50 4:40 7:50 10:40pm	
AMC Mercado 20	5.8 mi
10:10am 1:00 4:05 7:00 10:00pm	
AMC Cupertino Square 16	7.4 mi
10:00 1:00 4:00 7:00 10:00	

Web Images Places

translate I would like some local beer to Spanish

me gustaría un poco de cerveza local

"I would like some local beer" to Spanish

["I would like a beer" in Spanish? - HowDoYouSay.net](#)

[www.howdoyousay.net/...spanish/L\\_woul...](http://www.howdoyousay.net/...spanish/L_woul...)

How to say I **would like a beer** in **Spanish**. Includes **translation** from English and pronunciation. ... I **would like** to purchase **some** boots. If I wanted raw food, I would have gone to a sushi bar instead!

[How to say would you like a beer in Spanish?](#)

Web Images Places



Converse naturally with Google, on any platform  
Accelerate how you get information & get things done



# Challenges and Methods

It has to work everywhere, every time

How we're getting there.

---



2006

2011

TECHNOLOGY

INNOVATION, THE INTERNET, GADGETS, AND MORE.

APRIL 6 2011 4:36 PM

## Now You're Talking!

Google has developed speech-recognition technology that actually works.

By Farhad Manjoo

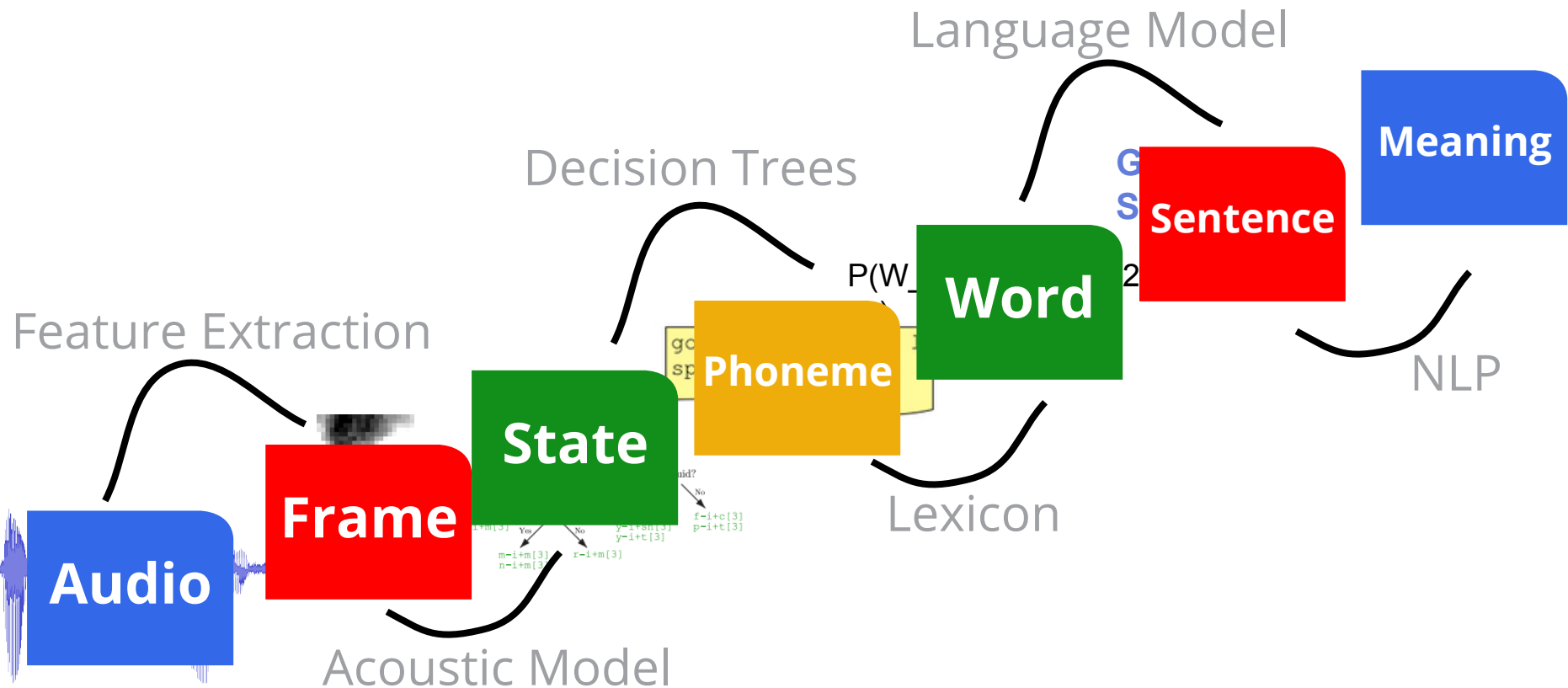


24

4

0





Audio Waveform



$$\operatorname{argmax}_W p(W | O) = \operatorname{argmax}_W p(O | W) p(W)$$

$W$



Word Sequence



Acoustic Model      Language Model

# Weighted Finite State Transducers (FSTs)

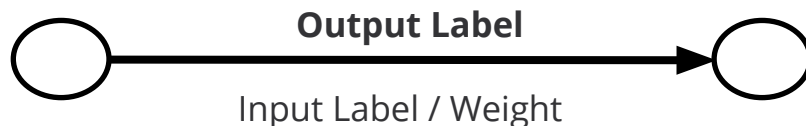
***Challenge: How to efficiently represent the very large phonetic, lexical, and language models needed for Voice Search and other LVR tasks.***

WFSTs one way to express this as probabilistic transductions.

A mathematically sound way to express probabilistic graphs and algorithms over them. (e.g. Viterbi, forward-backward)

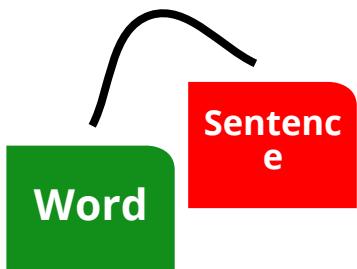
Powerful algorithms to combine and optimize these graphs.

Graphical representation:

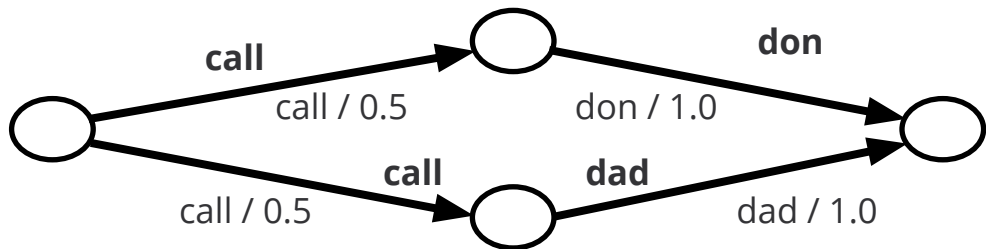


Mehryar Mohri, Fernando Pereira, and Michael Riley. "Weighted finite-state transducers in speech recognition." Handbook on Speech Processing and Speech Communication, Part E: Speech recognition, Springer-Verlag. 2008.

# Language Model



- Toy Example:
  - "call don"
  - "call dad"



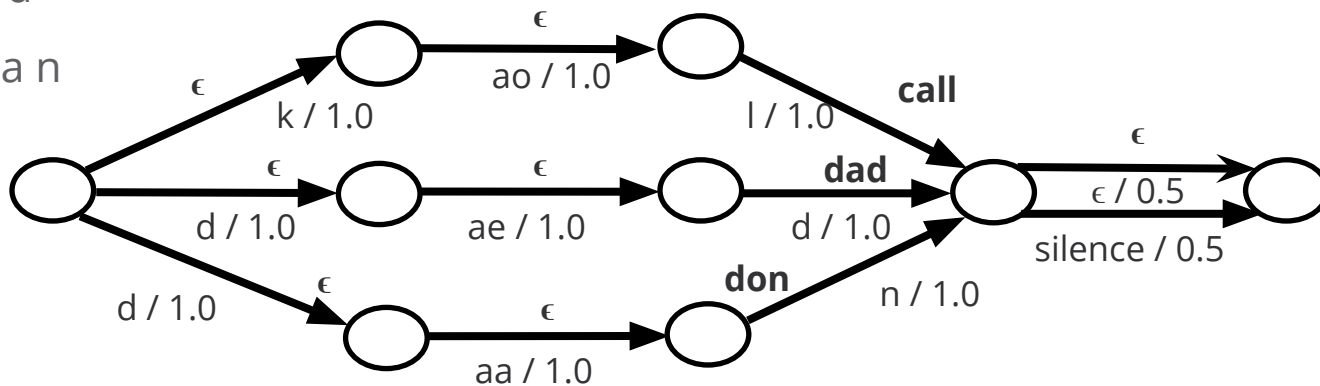
- In reality:
  - 100M+ states
  - estimated using billions of training examples.

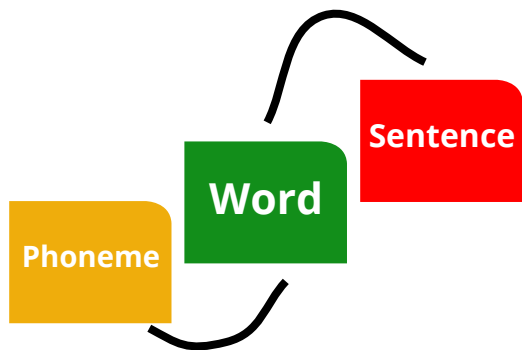
Phoneme **Word** Lexicon

call: k ao l

dad: d ae d

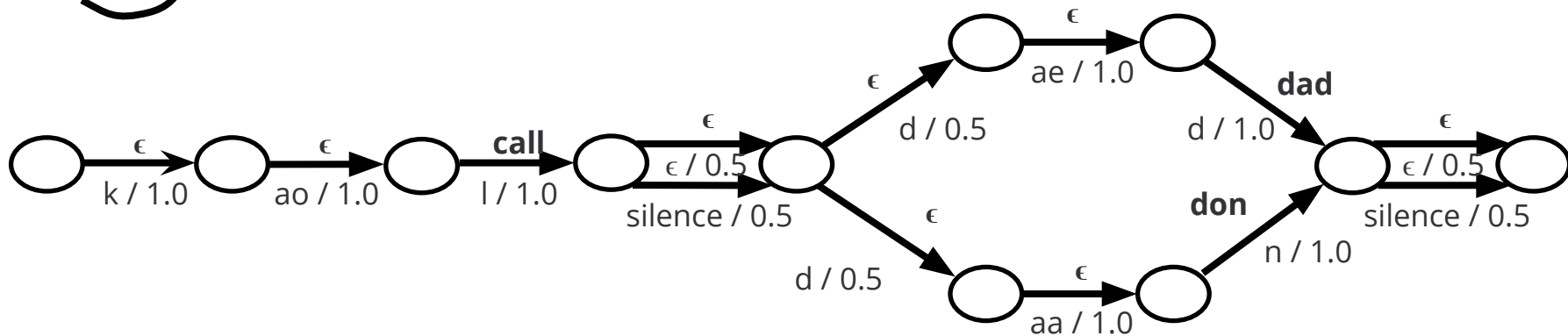
mom: d aa n

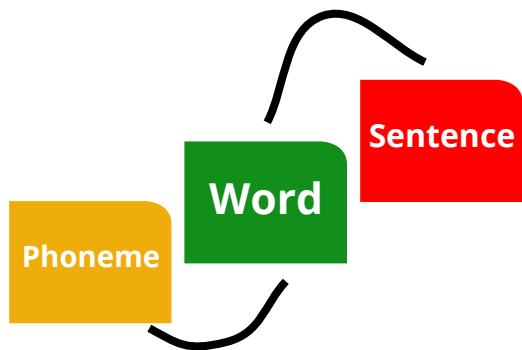




# Cascading Transductions: Composition

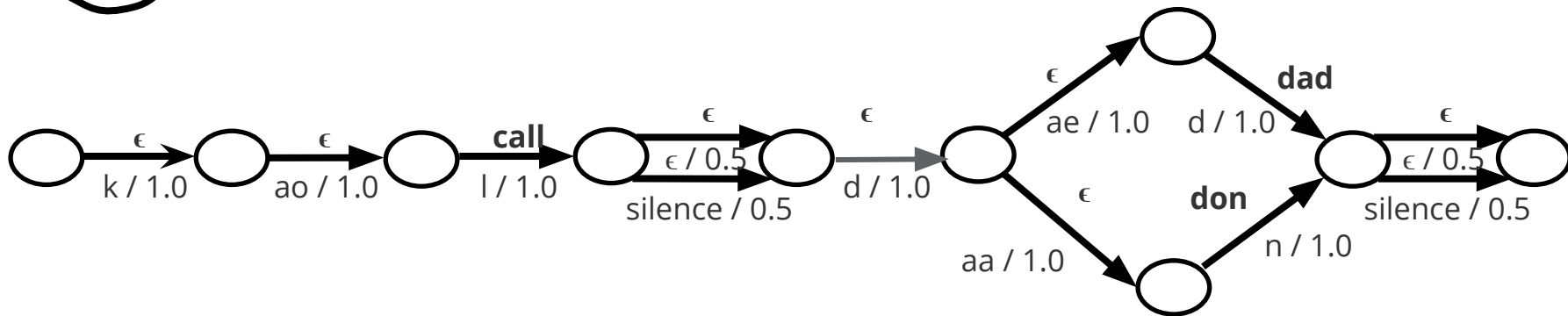
- Map *output* labels of Lexicon to *input* labels of Language Model.
- Join and optimize end-to-end graph.





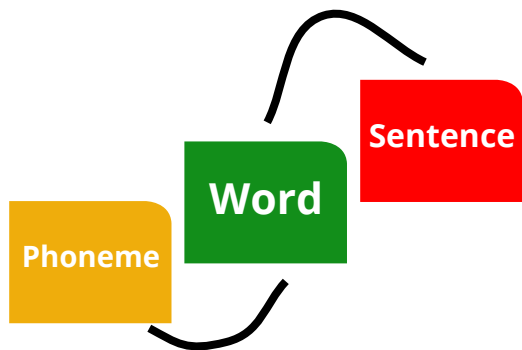
# Optimization via Determinization

- Redundancy causes excessive search
- Determinization creates an equivalent transducer that has no two transitions from a state with the same label



Other operations: Minimization, Epsilon removal, Weight pushing.

Uses our open-source OpenFst: [www.openfst.org](http://www.openfst.org)



# Context-Dependent Phonemes

- Acoustic models of phonemes built in **context** (typically triphonic): d\_aa\_n: /aa/ preceded by /d/ and followed by /n/
- Application of context-dependency to the decoder graph very naturally implemented by composition with a *context-dependency* finite-state transducer:

