Voice Search

– Information Access Via Voice Queries

Ye-Yi Wang
IEEE ASRU, Kyoto, Japan
December 10, 2007

Why the Topic of Voice Search

- It's an important speech understanding technology underlying many hot applications.

- It's a challenging problem and provides a fertile ground for research

  Typical automation rate 30%~60%

Research

Spoken Language Understanding

Form Filling (ATIS)

Show flights from Seattle to Boston on December 23

- ShowFlights
  - From=Seattle
  - To=Boston
  - Date=12/23

Call Routing (AT&T HMIHY)

My outlook does not show any new messages for about 5 hours

- Networking
- OS
- **Email**
- Hardware

Voice Search (Directory Assistance)

I need the number of Big 5

Sears Roebuck & Company	• 2200 148th Ave Redmond, WA • (425) 644-6526
Calabria Ristorante Italiano	• 132 Lake St S, Kirkland, WA • (425) 822-7350
Big 5 Sporting Goods	• 4315 University Way Seattle WA • (206) 547-2445

Spoken Language Understanding

	User input utterances		Target Semantics	
	Naturalness	*Input space*	*Resolution*	*Semantic space*
Form filling/ directed dialog	low	small	low	small
Form filling/ mixed initiative	low-med	small	high	small
Call routing	high	large	low	small
Voice search	med-high	large	low	med-large

Voice Search Applications

Residential DA
- French Telecom
- Bell Canada
- Telecom Italia
- Bellcore

Auto-attendant
- Phonetic Systems
- IBM
- AT&T
- Microsoft

Stock Quote
- Tellme

Business DA
- Tellme
- Nuance
- Jingle Networks
- Google
- Microsoft

Product Rating
- Microsoft

Music search
- Daimler
- Microsoft

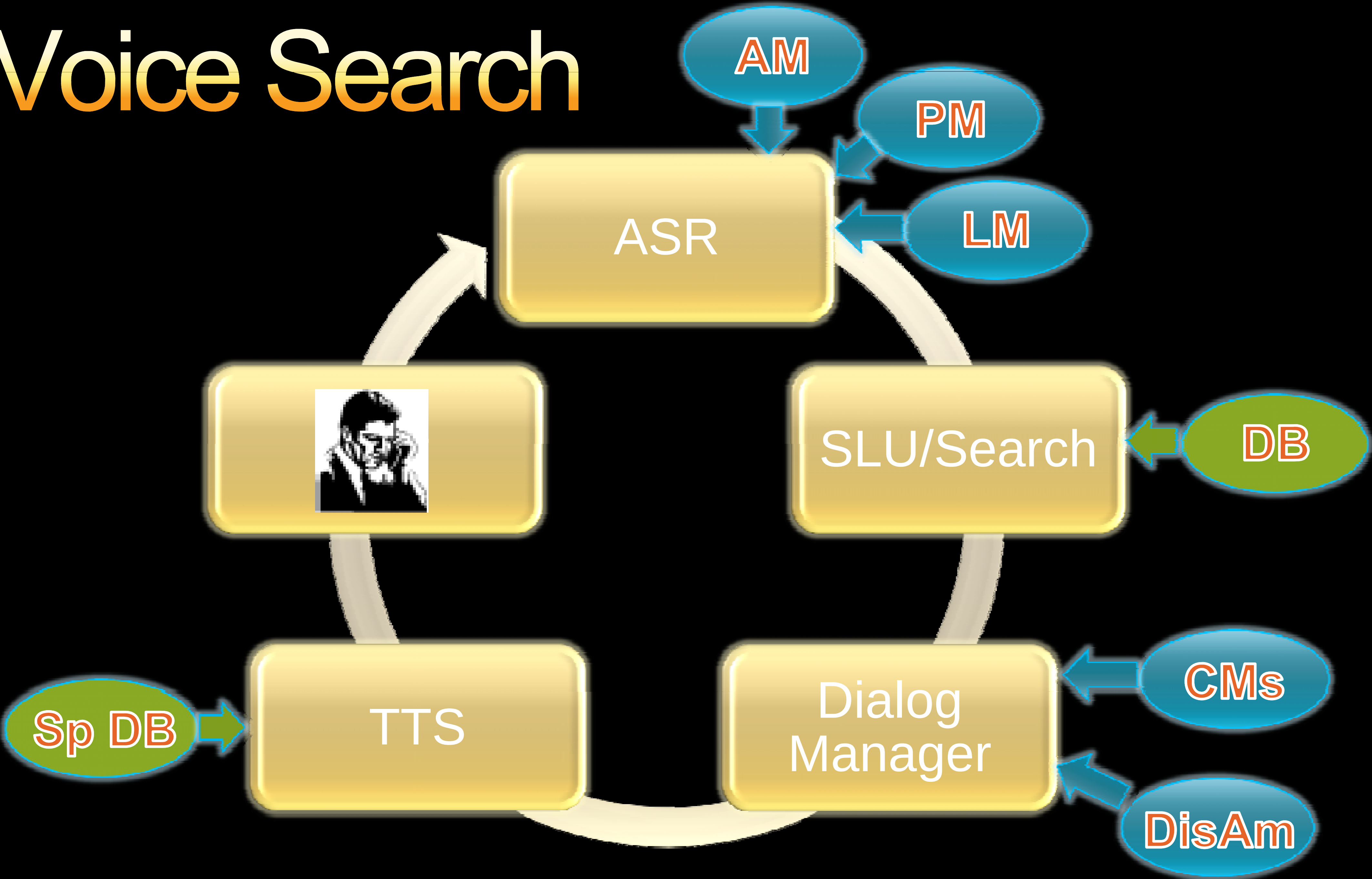
Conference papers
- Carnegie Mellon
- AT&T,ICSI, Edinburg Univ. etc

1995

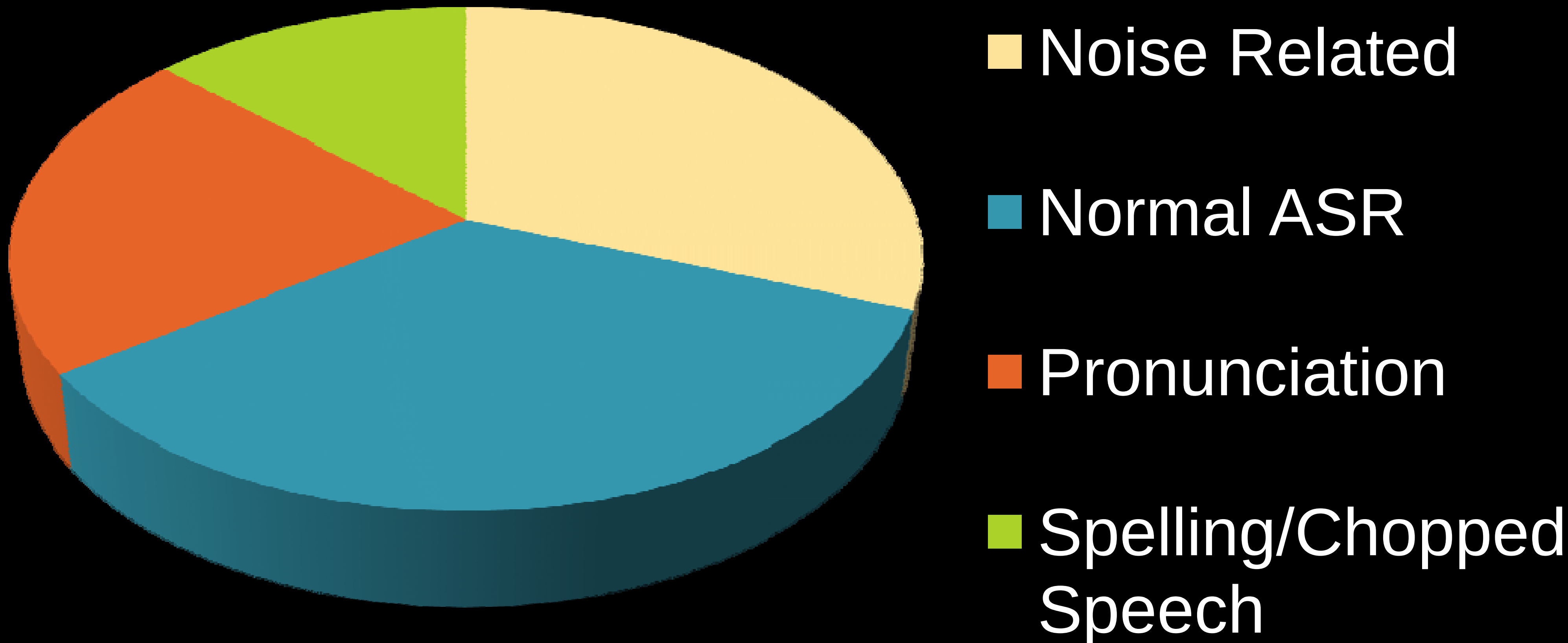
2007

Research

Voice Search



ASR in Voice Search



Y. Gao, *et al.* "Innovative approaches for large vocabulary name recognition." *ICASSP 2001*

Challenges (ASR)

Acoustic Model	• Noisy environment • Different channel conditions • Speaker variance
Pronunciation Model	• Foreign names • Unseen words • Pronunciation variance
Language Model	• Large vocabulary • Linguistic variance • Little training data from users

Research

Challenges

SLU/Search	• Huge semantic space • Linguistic variance
Dialog Management	• Multi-source uncertainties • High level confusability
TTS	• Large vocabulary • Foreign names • Unseen words
Feedback Loop	• System tuning is a necessity • High maintenance cost

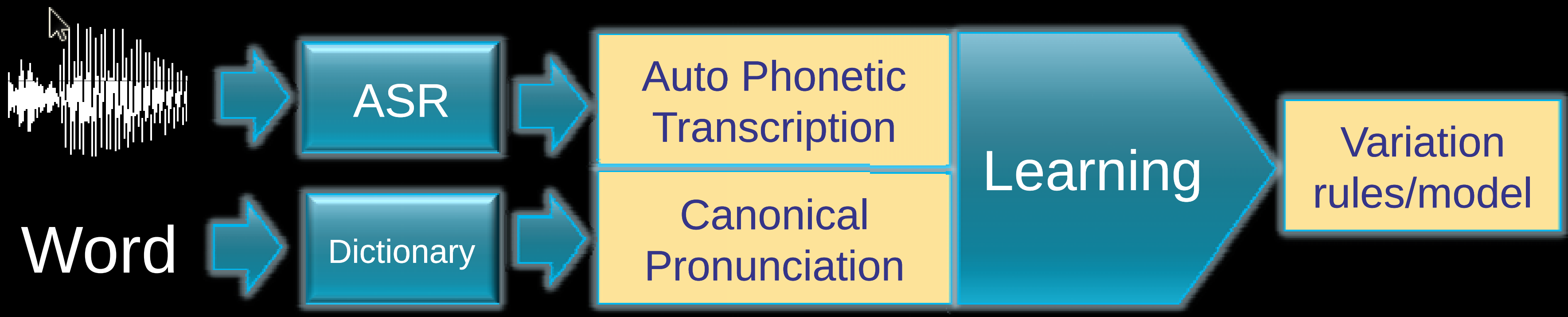
Acoustic Modeling

- Acoustic model clustering
  - More precise modeling of noisy environment and channel conditions
- Massive adaptation with most recent calls
  - More precise modeling of speaker variance
- Self-adaptation (2-pass online, unsupervised)
  - Better models for unseen speakers

Y. Gao, *et al.* "Innovative approaches for large vocabulary name recognition." *ICASSP 2001*

Pronunciation Modeling

- Pronunciation variants – Common Approach



Requires canonical pronunciation (No OOV)

- Derive pronunciation from acoustics only

B. Ramabhadtan, L. R. Bahl, P. V. deSouza, and M. Padmanabhan, "Acoustics-only based automatic phonetic baseform generation," *ICASSP* 1998

N-Best Rescoring with Pronunciation Distortation

$$W^* = \underset{W}{\operatorname{argmax}}\, p(W|A) = \underset{W}{\operatorname{argmax}} \sum_{\tau_w} p(W, \tau_w | A)$$

$$\approx \underset{W,\tau_w}{\operatorname{argmax}}\, p(W, \tau_w | A) \approx \underset{W,\tau_w}{\operatorname{argmax}}\, p(\tau_w)\, p(A|\tau_w) p(W|\tau_w)$$

$$p(\tau_w) = p(\eta_w \delta_w) = p(\delta_w | \eta_w) p(\eta_w)$$

$p(A|\tau_w)$: from acoustic model

$p(W|\tau_w)$: from data/knowledge source

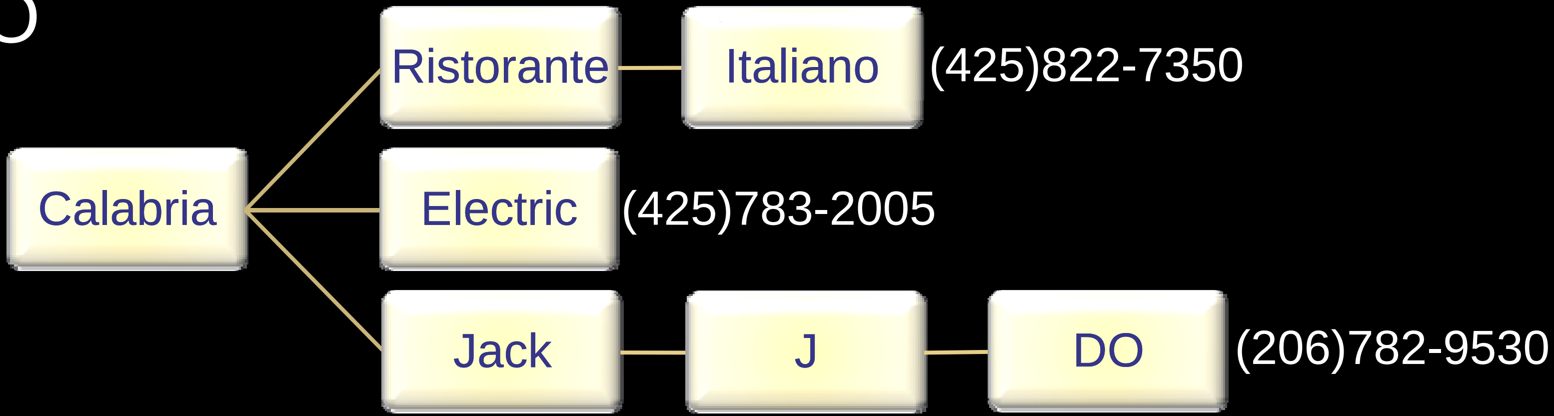
F. Béchet, R. De Mori, and G. Subsol, "Dynamic generation of proper name pronunciations for directory assistance," *ICASSP* 2002

Finite State Transducer LMs

Calabria Ristorante Italiano

Calabria Jack J DO

Calabria Electric



- DB listings don't match users' expressions
- Exact rule matching – lack of robustness
- High perplexity due to the lack of probabilities

Finite State Signature LMs

- *Signature*: Subsequence of a listing that uniquely identifies the listing
  - Listing 1: “3-L Hair World on North 3rd Street”
  - Listing 2: “Suzie’s Hair World on Main Street”
    - Signatures: “3-L”, “Hair 3rd”, and “Hair Main”
    - Non-Signatures: “Hair World” and “World on”
- FST LM:

S:= 3-L Hair World? On? North? 3rd ? Street? :1 | 3-L Hair? World? On? North 3rd ? Street? :1 |
3-L Hair? World on? North? 3rd ? Street? :1 | 3-L? Hair World? On? North 3rd ? Street? :1 |
Suzie’s? Hair World? On? Main Street? :2 | Suzie’s Hair World? On? Main? Street? :2 |
Suzie’s Hair? World on? Main? Street? :2 | Suzie’s? Hair? World on? Main Street? :2

Research

Finite State Signature LMs

- LM is constructed from listing database. No training data from users is required.
- Presumption: users will not skip the words that may lead to ambiguity
  - May not be true: users may not have the knowledge about confusable entries.
  - E.g. “Calabria” for “Calabria Ristorante Italiano” when “Calabria Electric” is in the DB

E. E. Jan, B. ı. Maison, L. Mangu, and G. Zweig, "Automatic Construction of Unique Signatures and Confusable Sets for Natural Language Directory Assistance Applications," *Eurospeech* 2003

Statistical N-gram LMs

- Robust (exact match not required)
- Well-studied smoothing algorithms
- Training data
  - BBN: user queries + FRN listings
  - Microsoft Research:

$$p(w) = \lambda p_t(w) + (1 - \lambda) p_l(w)$$

P. Natarajan, et al, "A scalable architecture for directory assistance automation," ICASSP 2002.
D. Yu, et al, "Automated Directory Assistance System - from Theory to Practice," *INTERSPEECH 2007*.

Research

Modeling Variations in LM

- For improving $p_l(w)$
- Words in a listing can be skipped. The probability for skipping a word is inversely proportional to its "importance."
- The "importance" of a word depends on its idf value and its position in the listing – initial words are more important.
- Each word has a transition probability to a the words related to the listing's category (e.g., restaurant, hospital)

SLU/Search

- Finite state transducer LM – not an issue
  - May be good enough for residential DA
  - Poor coverage, not robust for business DA
- Statistical n-gram language model
  - SLU/Search is required
  - Statistical model for robustness

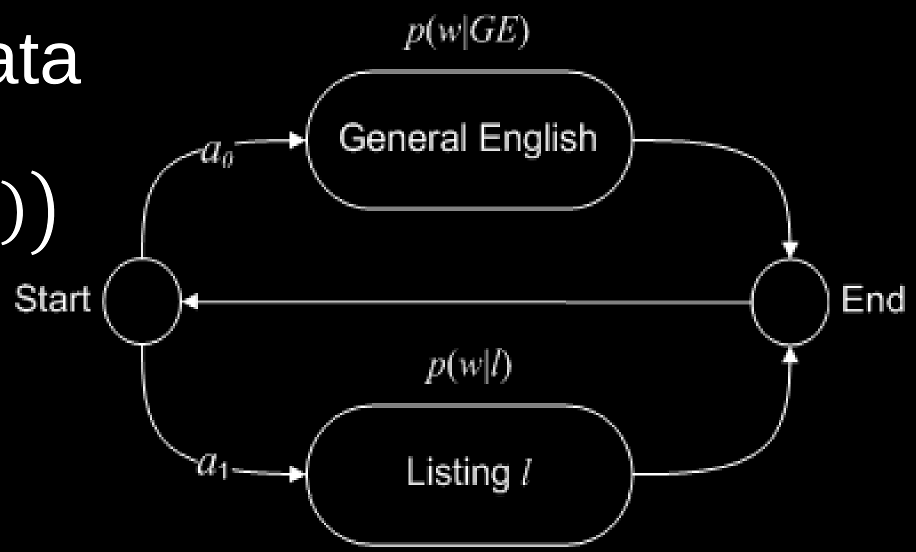
SLU/Search – Channel Model

$$\hat{L} = \underset{L}{\operatorname{argmax}}\, p(L|C,Q) = \underset{L}{\operatorname{argmax}}\, p(C,Q|L)p(L) = \underset{L}{\operatorname{argmax}}\, p(C|L)(Q|L)p(L)$$

$p(L)$: static rank

$p(C|L)$: directly estimated from data

$$p(Q|L) = \prod_{w \in Q} \left(a_0 p(w|GE) + a_1 p(w|L)\right)$$



Designed for Frequently requested names (FRNs)

P. Natarajan, R. Prasad, R. M. Schwartz, and J. Makhoul, "A scalable architecture for directory assistance automation," *ICASSP 2002*

SLU/Search – Tf*Idf VSM

- Tf*Idf weighted vector space model
  - Represent queries and listings as vectors
  - Each dimension represents the importance of a term (e.g., word, word bigram) in a query/document
  - The importance is proportional to the term's frequency in the query/document (TF). It reduces as the term occurs in many different documents – proportional to the logarithm of the inverse document frequency (IDF).
  - Measure the similarity as the cosine of the vector

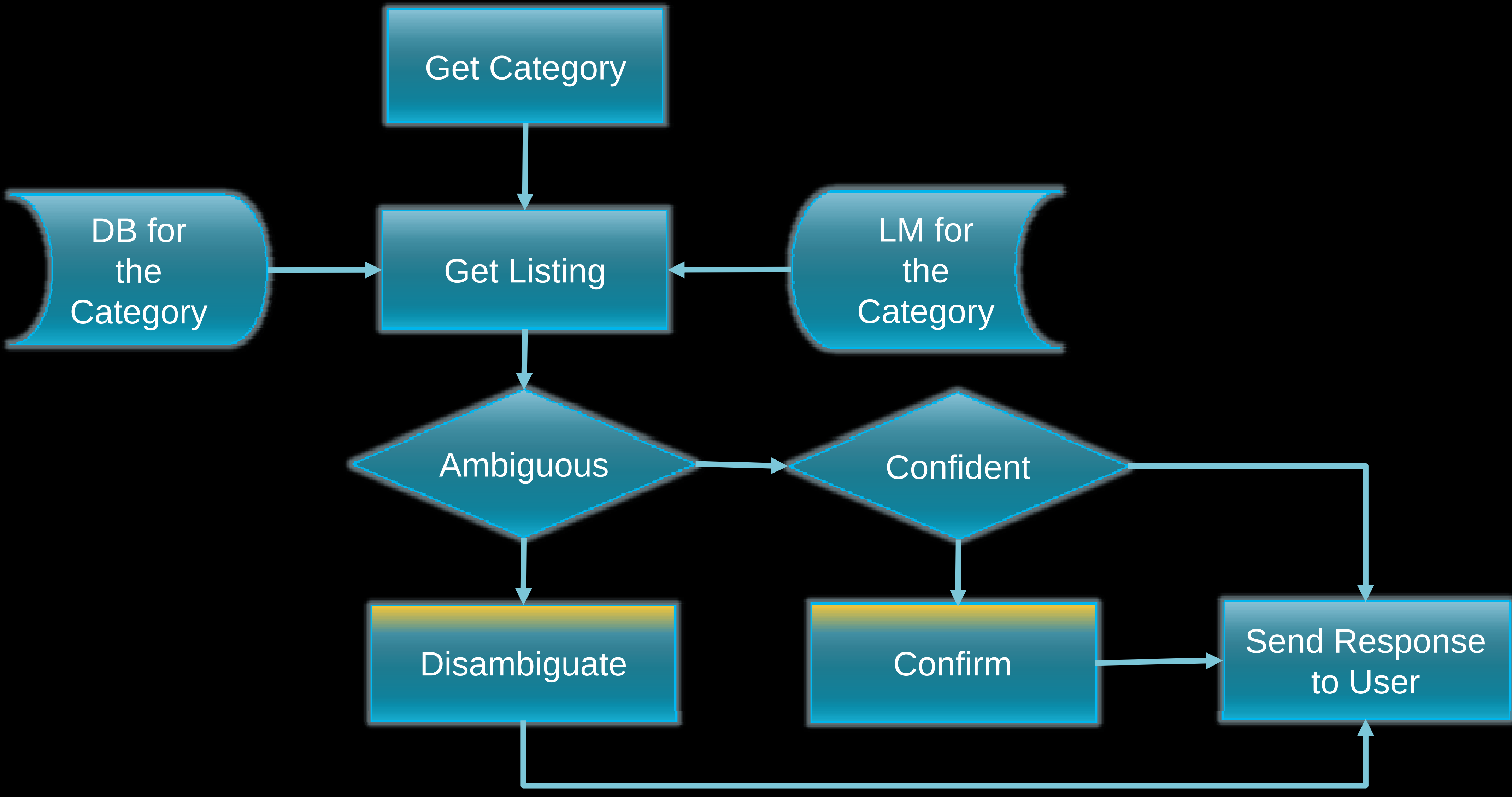
Research

SLU/Search – Tf*Idf VSM

- Enhancements
  - Duplicated words for short documents (listing) and queries – term frequencies are not reliable
    - “Big 5” will get “Big 5 Sorting Goods” instead of “5 star 5”
  - Category smoothing
    - “Calabria Ristorante Italiano” is favored over “Calabria Electric” on the query “Calabria Restaurant”
  - Character-ngrams as terms
    - Lime Wire vs. Dime Wired (ASR Error)
    - $Lim Lime ime_ me_W e_Wi _Wir Wire ire$
    - $Dim Dime ime_ me_W e_Wi _Wir Wire ired red$

Research

Dialog Management



Disambiguation Strategy I

System: Say the number of the item you are looking for, or say none of them

1. Star Wars Trilogy, the DVD
2. Star Wars Trilogy computer game

User: None of them

System: Let’s do that again:

1. Star Wars light saber
2. Star Wars Return of the Jedi

Research

Disambiguation Strategy II

- Say yes when you hear the item you want

  1. Star Wars Trilogy, the DVD
  2. Star Wars Trilogy computer game
  3. Star Wars light saber
  4. Star Wars Return of the Jedi
  5. ……

Research

Disambiguation Strategy III

System: Are you looking for a DVD, a game or a toy?

User: DVD.

System: I found several items. Say the number of the item you are looking for , or say none of them

1. Star Wars Trilogy, the DVD
2. Star Wars Return of the Jedi

Summarization in Disambiguation

User: *Tell me about restaurants in London.*

System: *I know of 596 restaurants in London. All price ranges are represented. Some of the cuisine options are Italian, British, European, and French.*

User: *I'm interested in Chinese food.*

System: *I know of 27 restaurants in London that serve Chinese cuisine. All price ranges are represented. Some are near the Leicester square tube station.*

User: *How about a cheap one?*

System: *I know of 14 inexpensive restaurants that serve Chinese cuisine. Some are near the Leicester Square tube station. Some are in Soho.*

J. Polifroni and M. Walker, "An Analysis of Automatic Content Selection Algorithms for Spoken Dialogue System Summaries," *SLT 2006*.

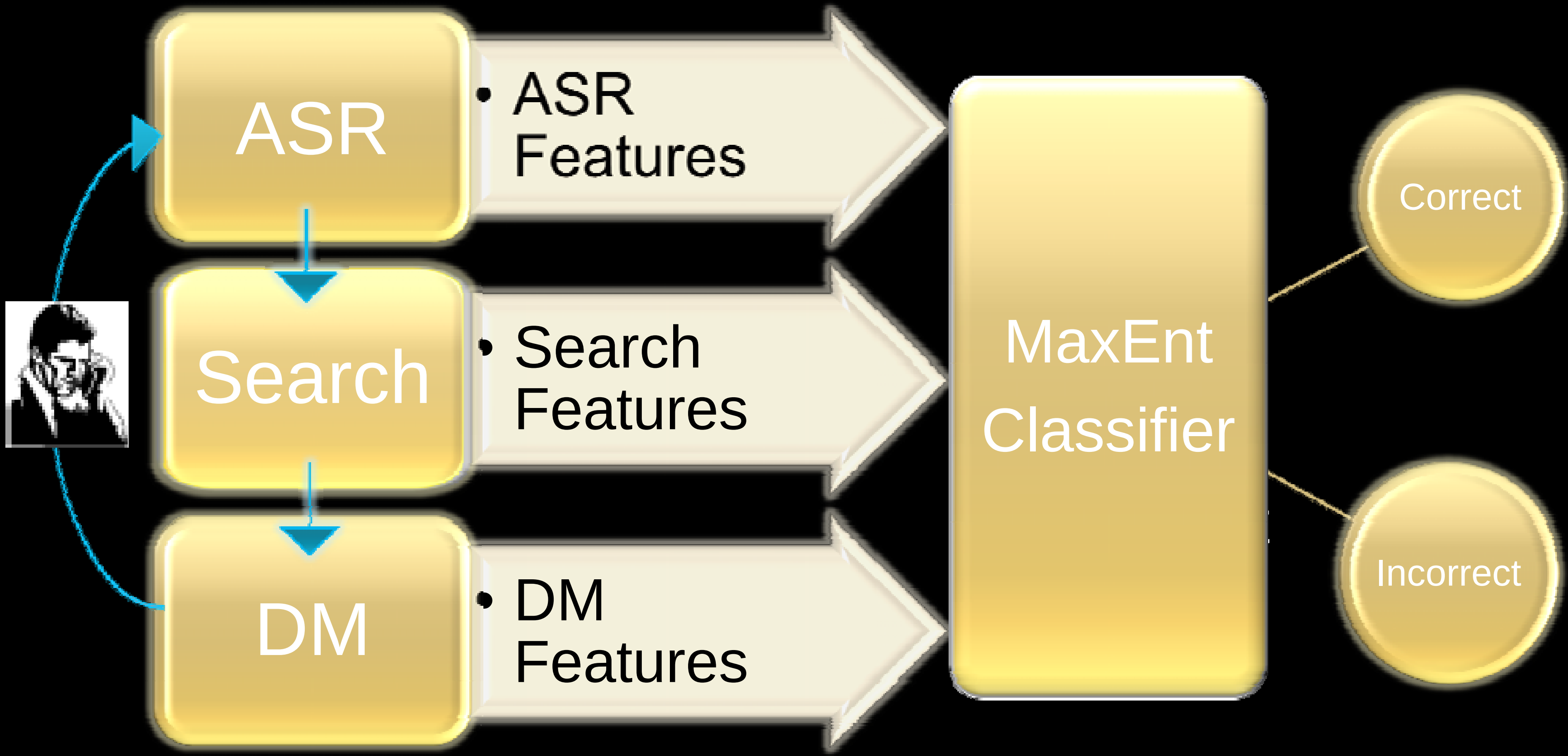
Research

Confidence Measures

- Early (residential) DA used ASR confidence.
- CMs must accommodate uncertainty from all components for other Voice Search app.
- Conditional probability P(corr | evidence).
- Statistical classifiers
  - BBN – Generalized Linear Classifier
    - Required and allowable word sets (for FRNs)
  - Microsoft Research: Maximum Entropy

P. Natarajan, R. Prasad, R. M. Schwartz, and J. Makhoul, "A scalable architecture for directory assistance automation," *ICASSP 2002*

Confidence Measures



Y.-Y. Wang, D. Yu, Y.-C. Ju, G. Zweig, and A. Acero, "Confidence Measures for Voice Search Applications," *INTERSPEECH* 2007.

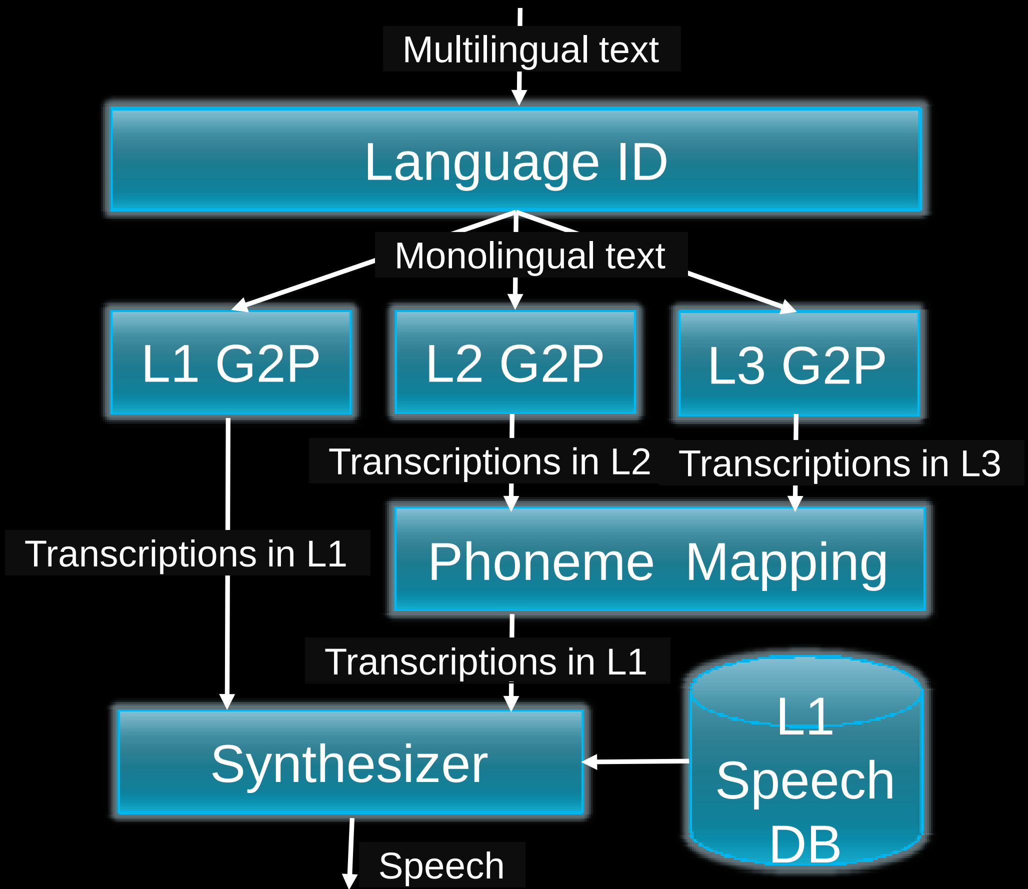
Research

Features for CMs

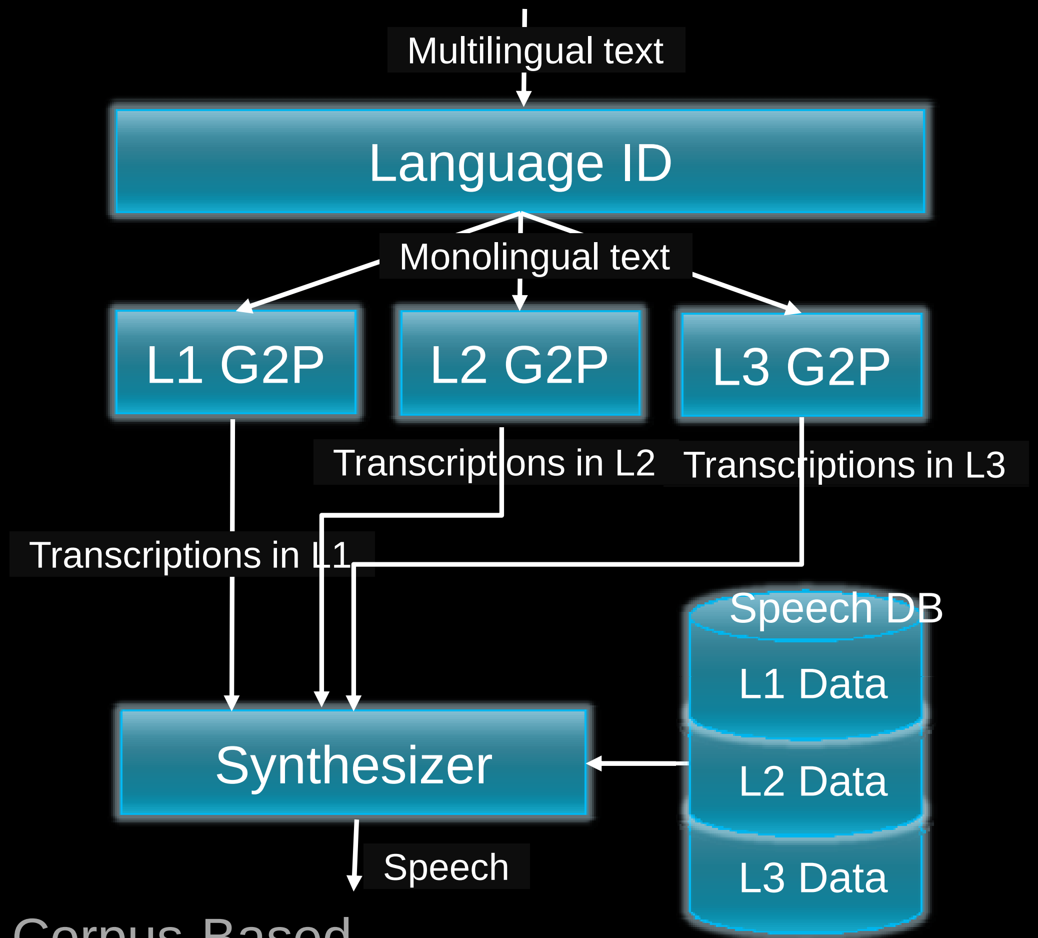
ASR	• ASR confidence score • ASR semantic confidence • [No ASR internal features]
Search	• TF*IDF VSM scores w/ and w/o category smoothing • VSM score gap from the best hypothesis • Normalized # of char. matches btw. ASR and hypo. • Covered/uncovered words' IDF ratio
DM	• Same hypothesis made in previous turn • Dialog turn • City match (application-specific)
Combined	• ASR confidence on the word with max. IDF value • Joint ASR confidence score/TF*IDF

Multilingual Text-to-Speech

- Phoneme Mapping



- Multilingual Speech DB



F. Deprez et al, “Introduction to Multilingual Corpus-Based

Learning in the Feedback Loop

- Identifying design/implementation flaws from log data after deployment
- Telecom Italia: identifying linguistic variants
  - Phonetic transcription
  - Furthest distance clustering of transcripts
    - Distance: phone specific ins-del-sub cost
  - Central element of a cluster used as a linguistic variant

C. Popovici, et al, “Learning new user formulations in automatic directory assistance,” ICASSP 2002

Research

Learning in the Feedback Loop

- Microsoft: finding uncovered semantics

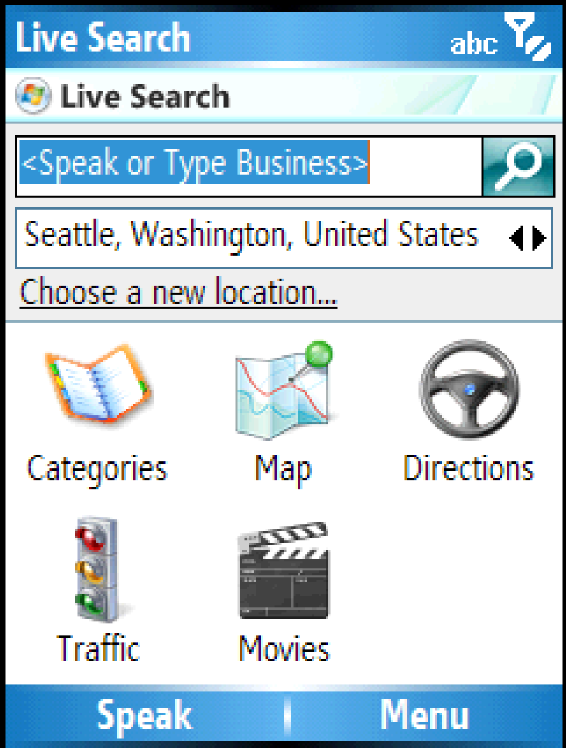
  e.g. auto-attendant does not cover "security," "shuttle service," "receptionist in building 99."

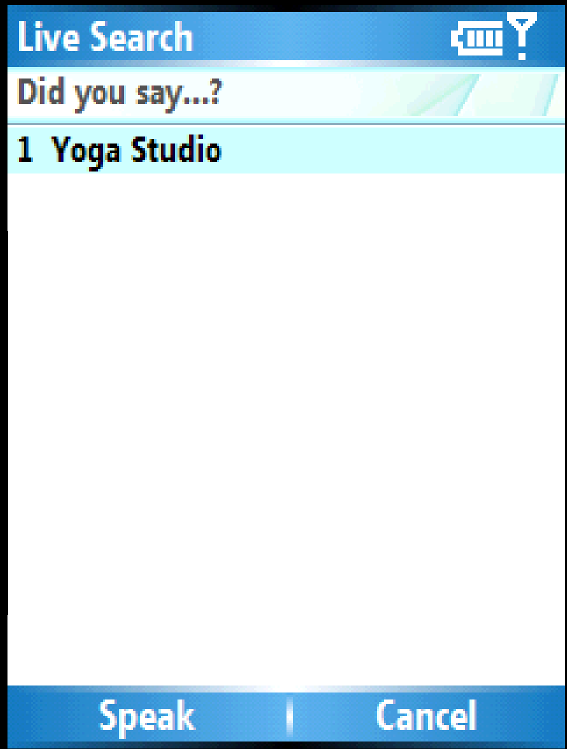
- PLSA-like clustering:

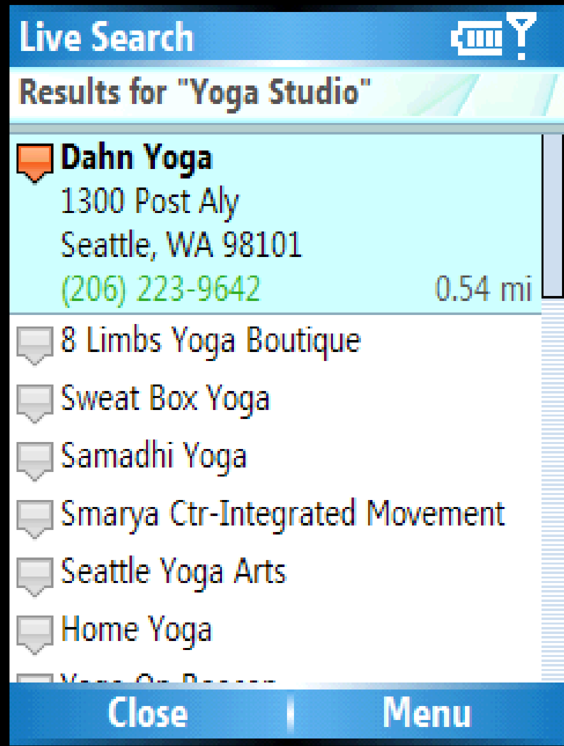
$$p(x) = \sum_{c,w} p(x, w, c) = \sum_{c,w} p(x|w)p(w|c)p(c)$$

X. Li, et al, "Unsupervised Semantic Intent Discovery from Call Log Acoustics," ICASSP 2005

Multimodal Voice Search







- *Two sweet features added... gas and voice... this software rocks!*
- *Isn't this program great! I feel like the NSA guys in enemy of the state*
- *Hey Google map user put up your stylus and check out Live Search. No stylus needed. Now that's handy or should I say one-handed.*

Summary

- Voice Search is an important type of spoken dialog that has been a hot topic recently.
- Challenges are from all components of SDSs: AM, PM, LM, SLU, DM, TTS.
- Some of solutions are reviewed.
- Voice Search remains a fertile research field.

The content of this talk will appear in the upcoming Special Issue on SLT of the IEEE Signal Processing Magazine.