A New Collaborative ASR Paradigm: Is the Glass Half Full or Half Empty?

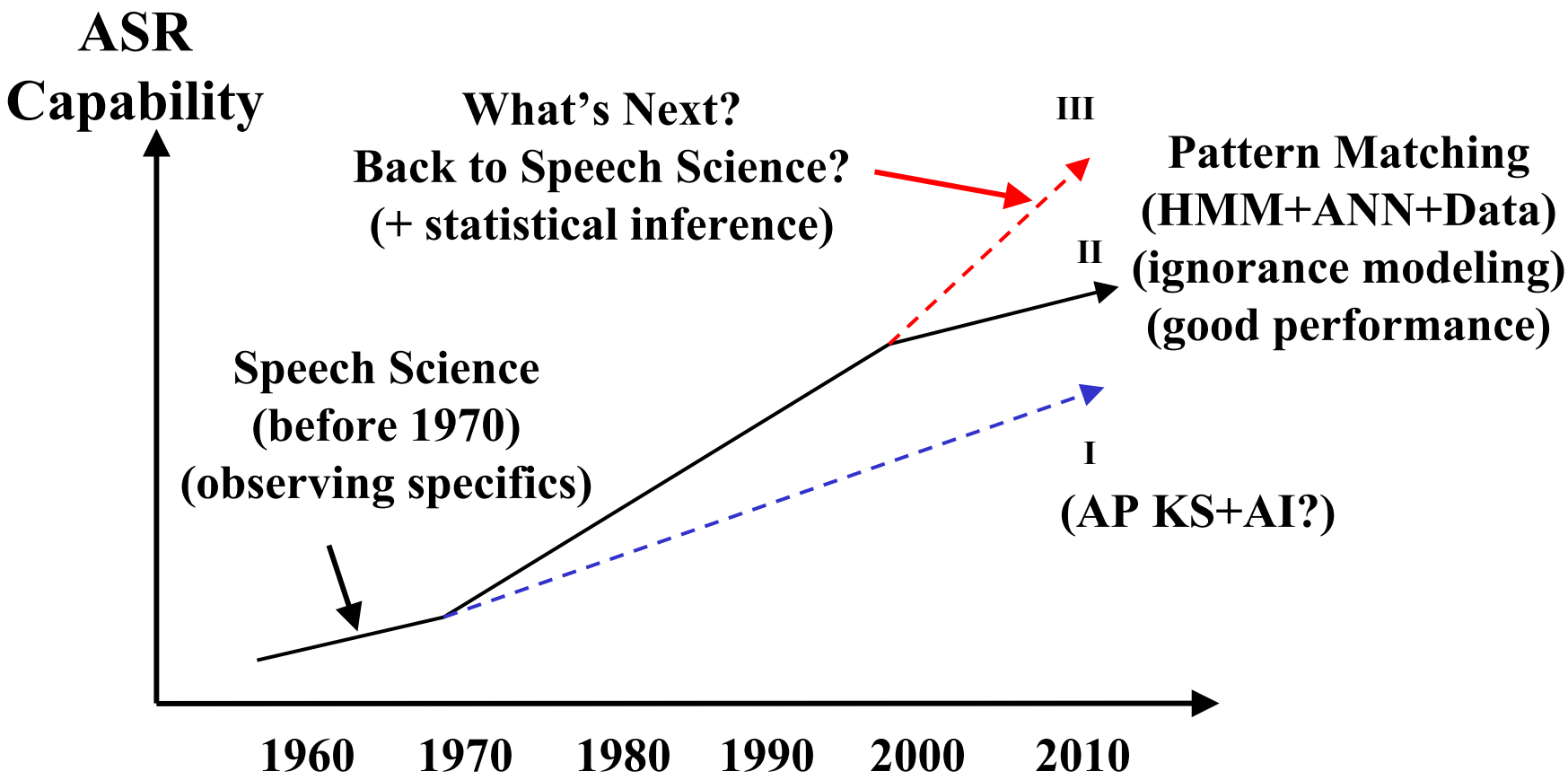
Chin-Hui Lee
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
chl@ece.gatech.edu

NSF Symposium on NG ASR at Georgia Tech, October 7, 2003

Outline

- **ASR Approaches: a historic perspective**
  - from top-down decoding to bottom-up detection and verification
  - from KS-ignorant to KS-intensive modeling
- **ASR Capabilities**
  - patter matching paradigm with data-driven modeling and DP
- **ASR Limitations**
  - robustness issues, gaps between ASR and HSR performance
- **A few examples of knowledge-based, data-driven ASR**
  - mimicking "foreign ears" with key-phrase detection, knowledge-based events for LVCSR, signal-dependent feature extraction
- **Future Work: Next-Generation ASR**
  - building a collaborative ASR community of the 21st century

ASR Technology Progress: "S" Curve



ASR: A Historic Perspective & Summary

I. Acoustic-Phonetic and AI Approaches to ASR

- distinctive features difficult to detect, non-invariant, not robust, non-deterministic; objective evaluation needed, a bottleneck
- plenty done in the ARPA SUR project, abundant literature

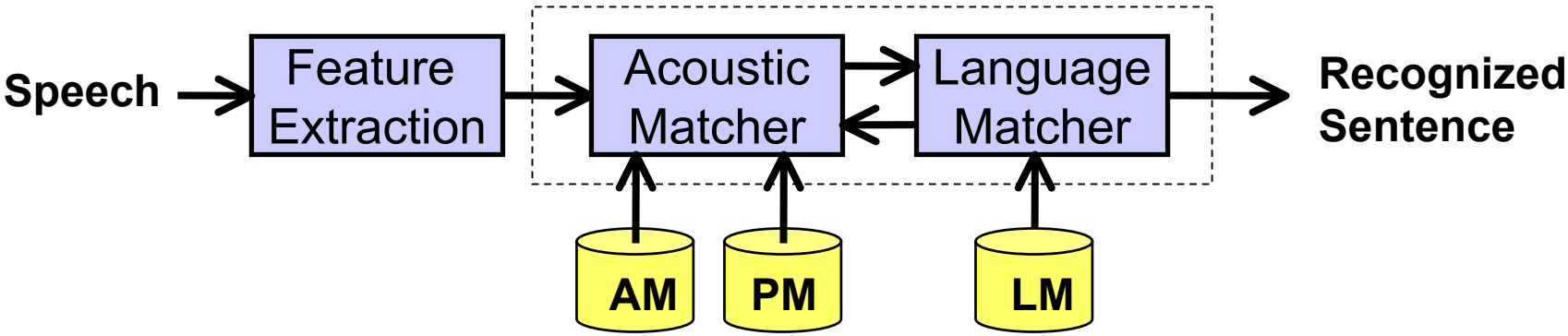
II. Pattern Matching Approaches to ASR

- data-driven, knowledge-ignorant, well-defined modeling
- difficult to integrate diverse and incomplete knowledge sources

III. Knowledge-Based, Data Driven Approach to ASR

- mimicking HSR by integrating KS into auditory perception
- making use of statistical modeling tools, data, DP, KS and DSP
- taking shape in mid-90's at BL; requiring community buy-in and an open platform for worldwide collaborative research

ASR Capabilities
(Knowledge-Ignorance Modeling)



- **Use powerful data-driven modeling tools, e.g. HMM, ANN, DP**
- **Do not rely on detailed speech and language specifications**
- **Give high performance in clean conditions for many small, medium and large vocabulary tasks in many languages**
- **Deployed in many applications and services**

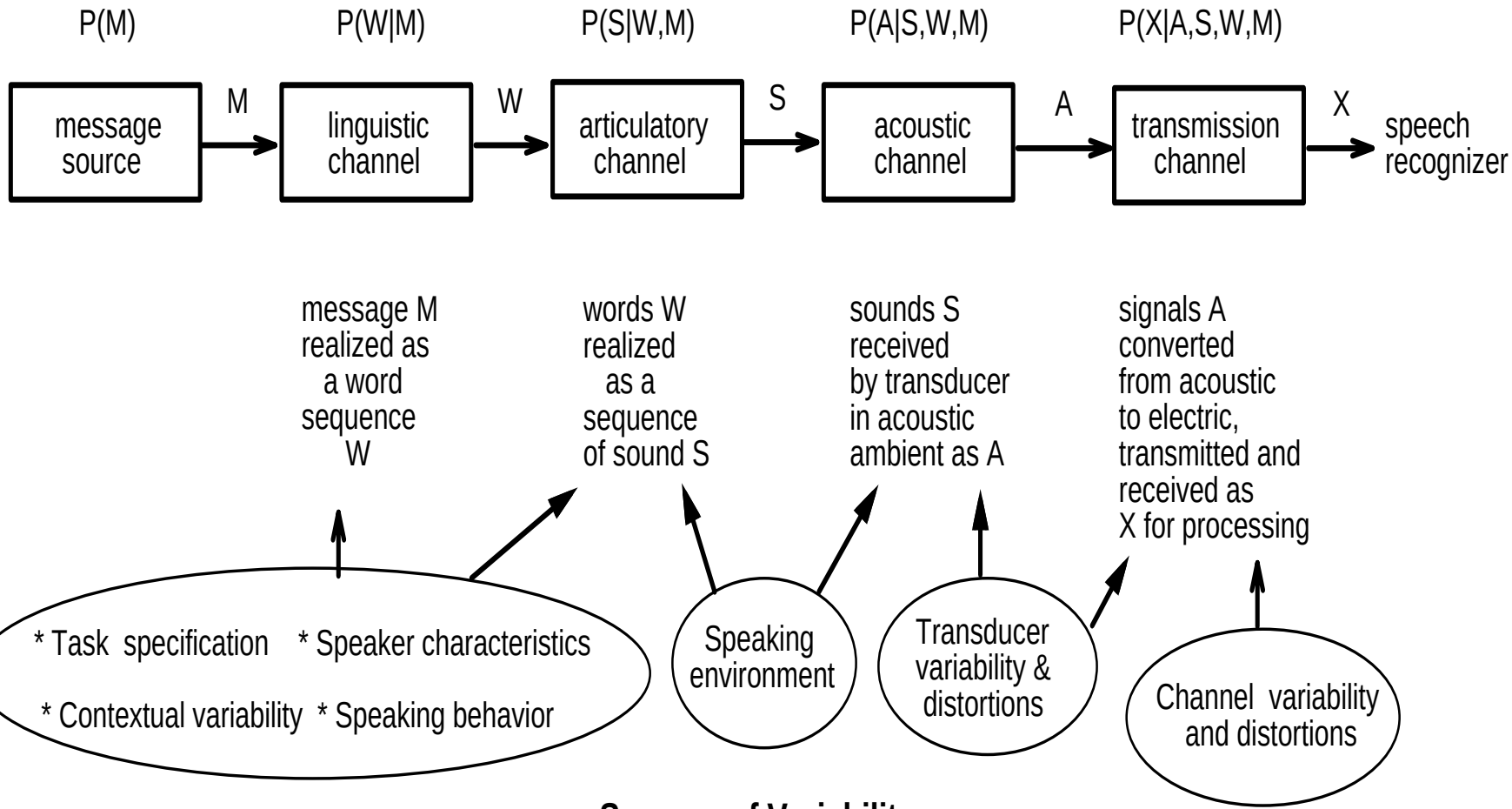
ASR Limitations

- **Integrate only a few knowledge sources: too task specific**
  - score normalization and synchronization of frame rate
  - missing channel characterization (but requiring complete KS specifications)
  - ill-formed utterances (OOV, OOG, OOT), need to follow specific protocols?
- **Miss collaboration opportunities**
  - not fully taking advantage of vast scientific knowledge in speech science
  - high entry barriers for small groups, need an open and common platform
- **Provide little diagnostic information (no KS constraints)**
  - facing cross-condition robustness but offering no fixes (major research focus)
  - asking for more data to design complex models (only incrementally better)
- **Give 10-100 times more errors than HSR in some cases (Why?)**
  - ASR and ASU are AI-complete problems?

➡ **How do we go beyond the current limitations?**

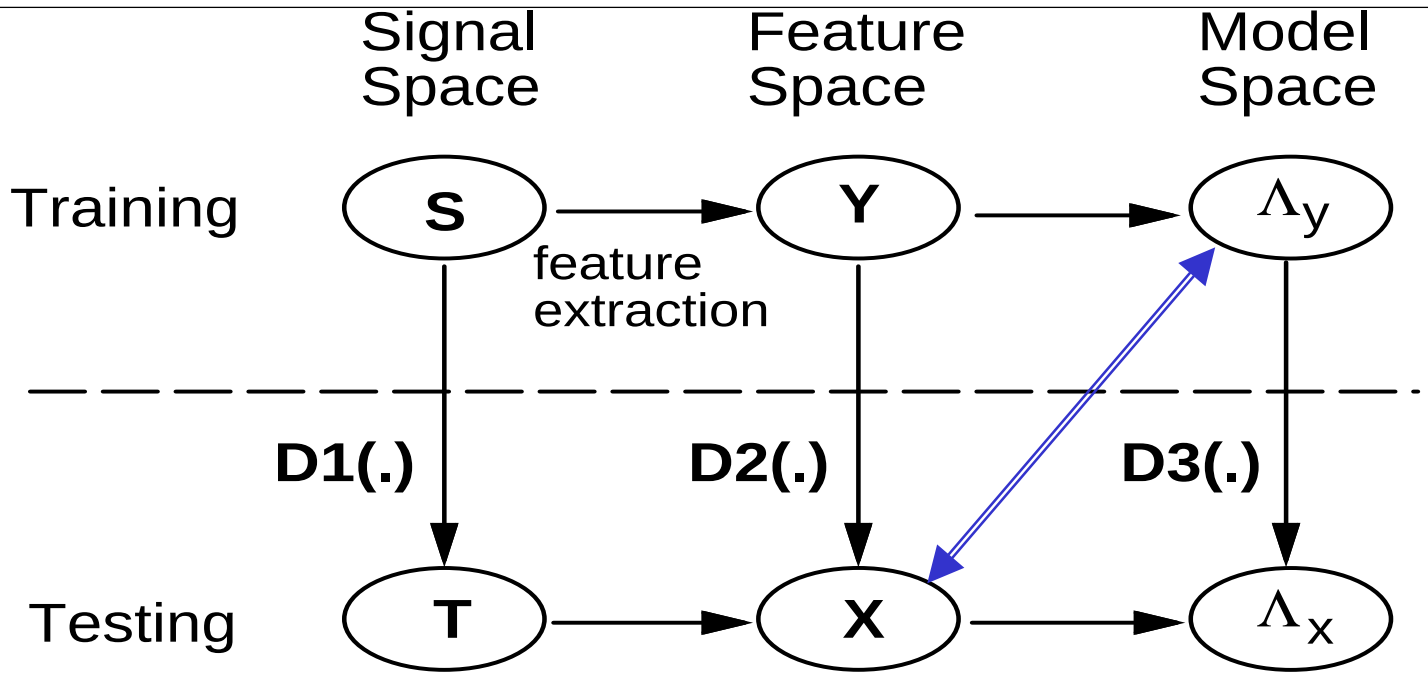
- incorporating knowledge sources into all modules of ASR system design !!

ASR: Complete Channel Characterization



Sources of Variability

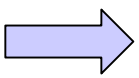
ASR Robustness: Incomplete Specifications



- **Mismatch between training and testing: main source of errors**
- **Main research topic, plenty of work, no solutions yet!**
- **Many knowledge source are difficult to characterize/integrate**

Motivations for Next Generation ASR

- **Address ASR Limitations (robustness issues)**
  - degradation under incomplete knowledge specifications
- **Enhance ASR Capabilities (fast progress again)**
  - ASR/HSR gaps, knowing where to improve, more literature
- **Lower ASR Entry Barriers (divide and conquer)**
  - enjoying contributions from big and small groups in all areas
- **Provide Broad ASR Collaboration Opportunities**
  - fully taking advantage of vast scientific knowledge in speech science, acoustics, linguistics, cognitive science, and others
- **Link ASR with all relevant research topics**
  - linking with speech production, auditory perception and neural processing



Knowledge-Based, Data-Driven ASR Paradigm

A Few Hints

- **ASR system as a not-so-good "foreign hear"**
  - spotting keywords in fluent speech is not hard for HSR and ASR (Kawahara/Lee/Juang, *T-SAP98*), how about other meta events?
- **Two-stage "teen-ty" and "e-set" discriminators**
  - second-pass detailed classification (e.g. nasal and stop detectors)
- **Knowledge-based features for LVCSR (WSJ)**
- **Event confidence (ASR ⇨ event verification)**
- **Learning from reading spectrograms and HSR**
  - bottom up integration of all knowledge sources, big or small

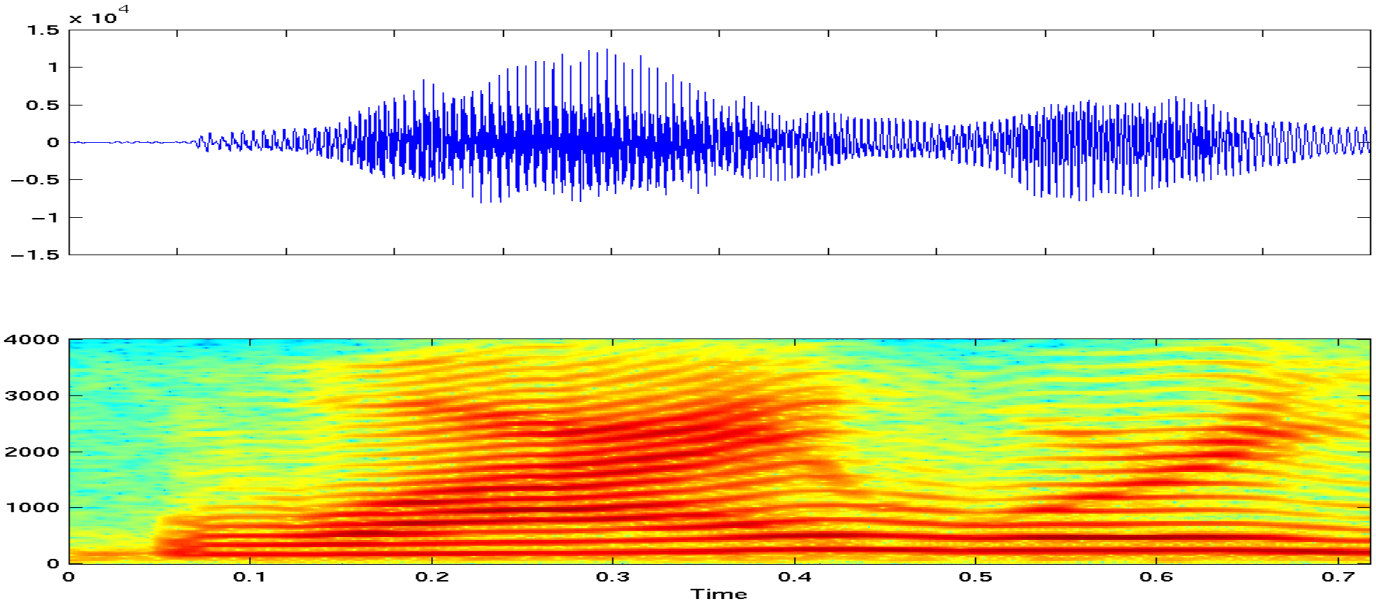
⇨ **Strong Message: bottom-up event detection**

  - but also taking advantage of data-driven modeling techniques

Center of Signal and Image Processing
Georgia Institute of Technology
CSIP

A Conceptual Speech Event Lattice

(Early and Meaningful Acoustic to Linguistic Mapping)

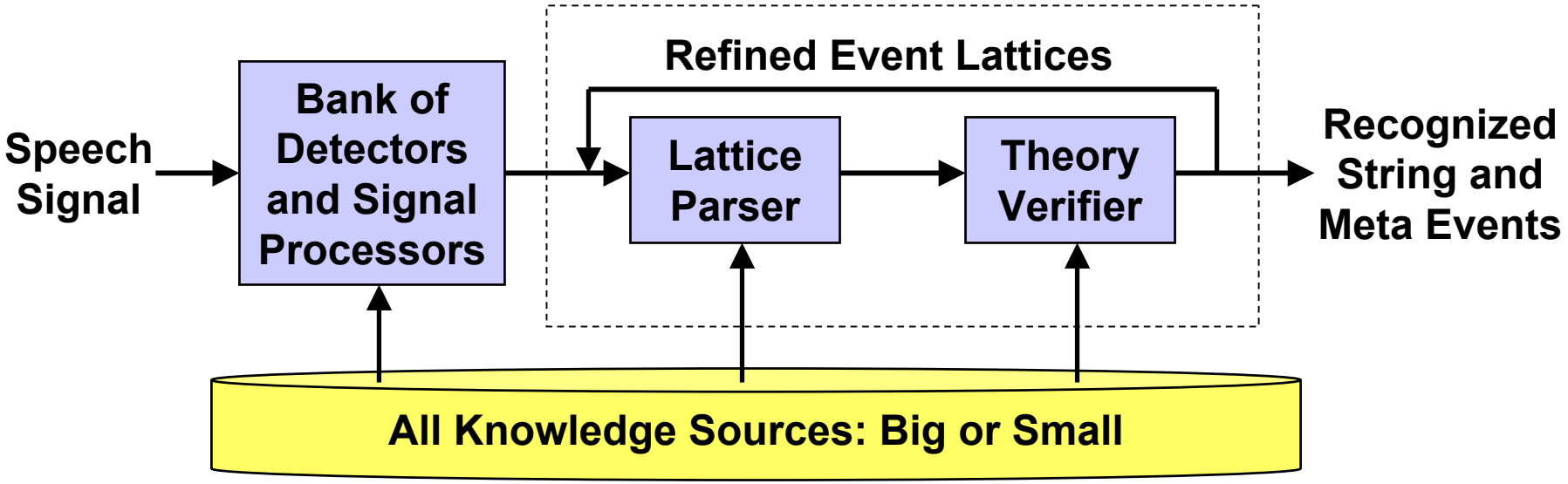


Voicing

Nasality

Frication

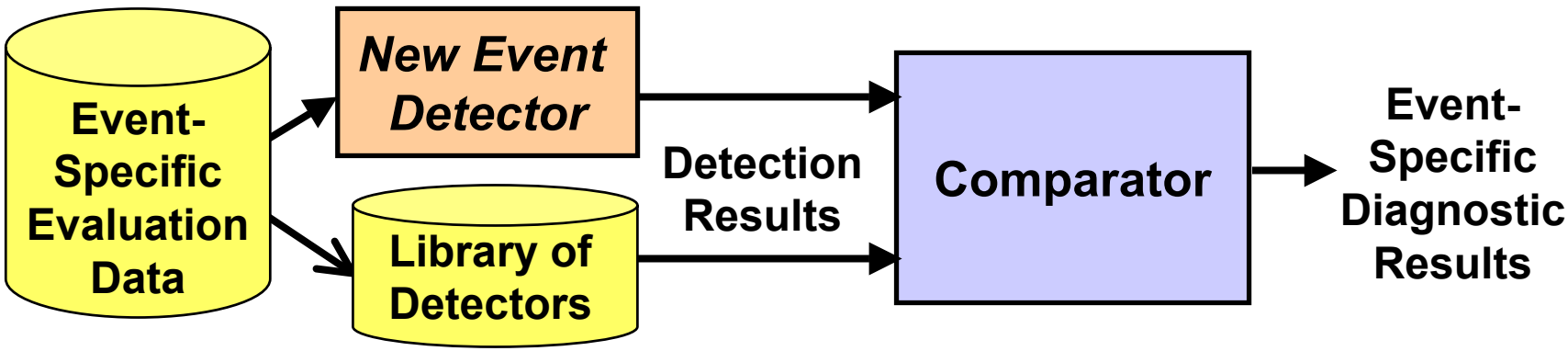
Proposed ASR Paradigm & Open Platform



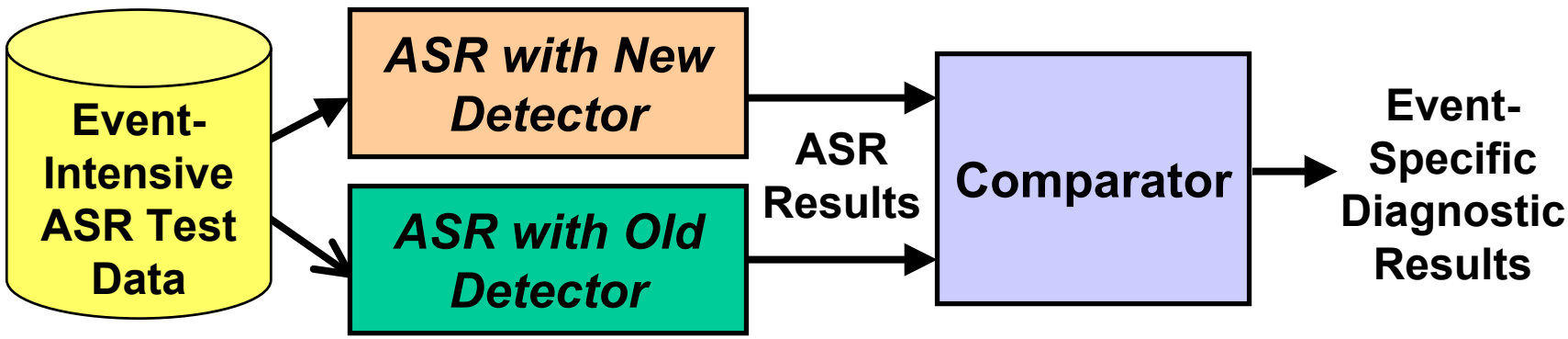
- No need for complete KS specifications (plenty of redundancy)
- "Optimal" components easier to design than the whole system
- A plug-'n'-play open platform for collaborative research
- Plenty of integration speech science and processing

A Proposed Evaluation Paradigm

1. Diagnostic Evaluation of Detectors (at event level):



2. Diagnostic Evaluation of Speech Recognizers:



Some Key Properties

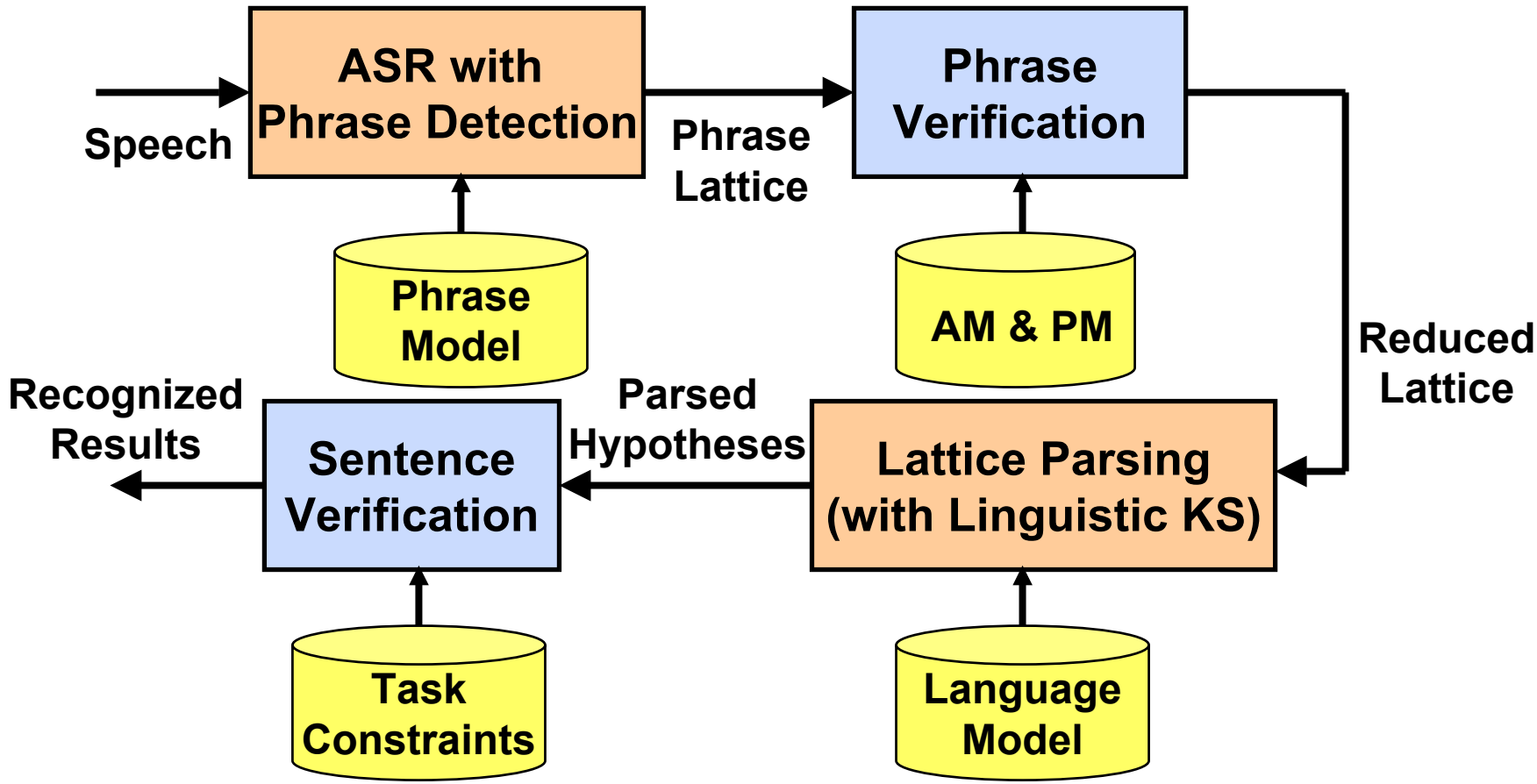
- **Robustness solution by "divide and conquer"**
  - **simulate and perfect robust single event detectors**
- **Sound-specific event detection & discrimination**
  - **e.g. 100 Ph.D.'s each contributing a best event detector**
  - **solving acoustic problems once and for all, no retraining**
  - **mixed analog and digital features, why fixed frame rate?**
- **Is blind speech data collection the solution?**
  - **how much more data are needed to solve the problem?**
  - **how about only collecting right data at the right time?**
- **Language-independent bank of event detectors**
  - **e.g. same nasal detectors for all languages**
  - **language-dependent event detectors are also critical**

Some Key Properties (Continued)

- **No model retraining needed, retaining only the best detectors, not the partial models**
  - **all modeling and verification tools can still be used**
- **Objective evaluation at all event levels**
  - **shared platform, plug-'n'-play modules and results**
- **Collaborative ASR Community**
  - **lowering ASR entry barriers, everyone can help**
- **Learning from speech science and processing**
  - **building upon the state-of-the-art and advancing fast**
- **Additive performance at all event levels**
  - **meaningful diagnostic information for improvement**

Flexible ASRU: Key Phrase Detection + Verification

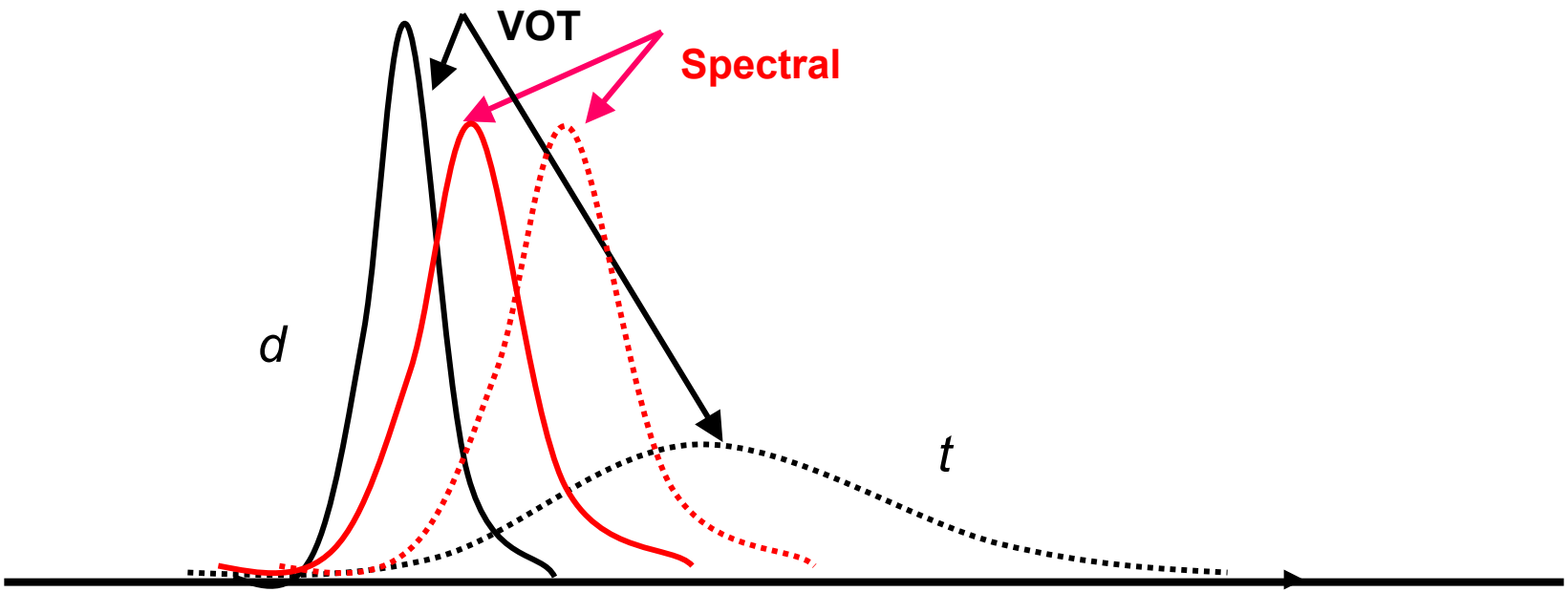
(Mimicking "Foreign Ears": Kawahara/Lee/Juang, *T-SAP*, 1998)



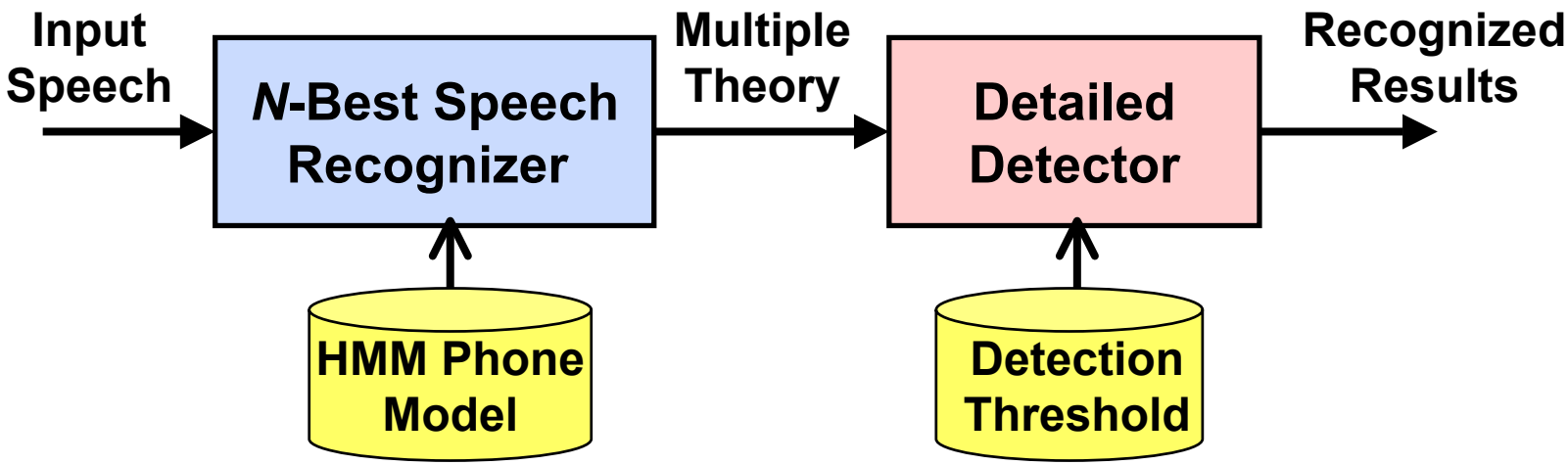
Greatly improve semantic accuracy for ill-formed utterances

Discriminative Features Improve Separation

- **VOT discriminates stops (Ramesh & Niyogi, *ICSLP98*)**
  - reducing E-set recognition error by 50% with two-stage processing
- **Sound-specific features & detectors (matched filters?)**
  - but an objective and rigorous evaluation paradigm is needed
- **Discriminative verification further enhances separation**



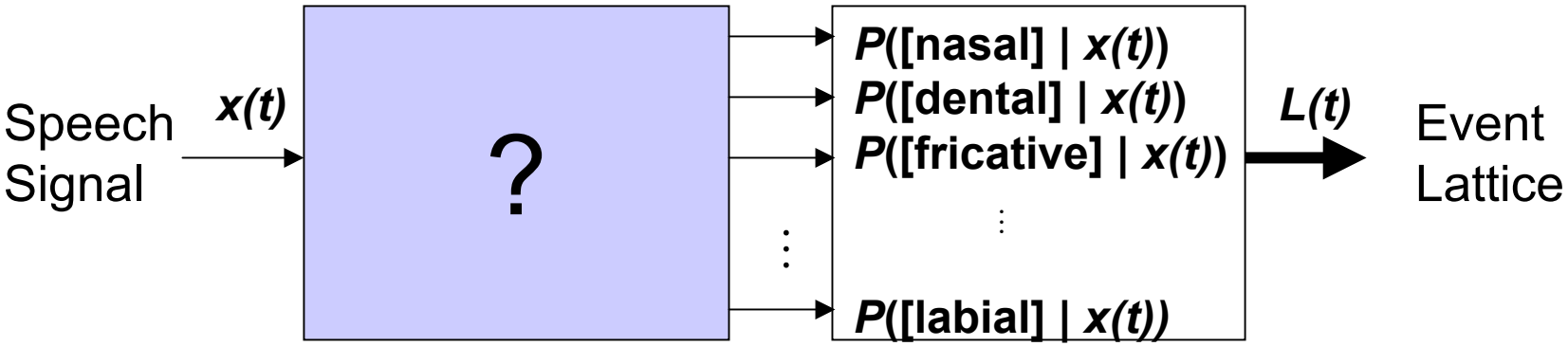
Two-Stage Speech Recognition
(Another Proof of Sound-Specific Detectors)



Two Straightforward Examples:

1. **ty-teen discrimination (nasal detectors): 45% less errors**
2. **e-set recognition (stop detectors): 50% less errors**

Detection of Distinctive "Events"



- **What kind of detectors should we use ?**
  - non-linear event detectors (activity levels & matched filters)
  - training could be performed from a set of labeled data
  - output can be interpreted as probability or confidence measure
  - features/events good for training classifiers and verifiers

Multilayer Perceptron (tools already available, other "best" detectors could also be designed, a major community effort)

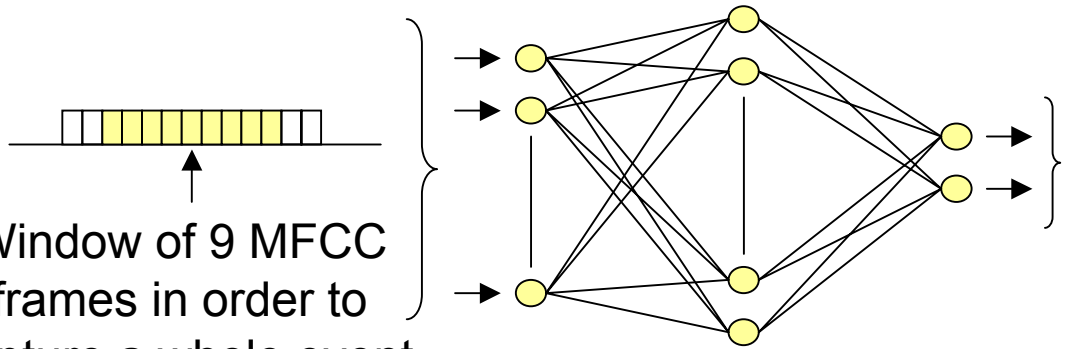
A Set of 60 Phone Attributes (for Clustering)

General	Vowels		Consonants		
Stop	Front Vowel	Rounded	Unvoiced	Non Anterior	Central Stop
Nasal	Central Vowel	Un-rounded	Voiced	Continuant	Back Stop
Fricative	Back Vowel	Reduced	Front Consonant	Non-Continuant	Voiced Fricative
Liquid	Long	I-Vowel	Central Consonant	Positive Strident	Unvoiced Fricative
Vowel	Short	E-Vowel	Back Consonant	Negative Strident	Front Fricative
Front	Diphthong	A-Vowel	Fortis	Neural Strident	Central Fricative
Central	Front Start	O-Vowel	Lenis	Syllabic	Back Fricative
Back	Fronting	U-Vowel	Neither Fortis or Lenis	Voiced	Affricate
	High		Coronal	Unvoiced	Not Affricate
Noise	Medium		Non Coronal	Stop	
Silence	Low		Anterior	Front Stop	

(J. Odell, Ph.D. Thesis, Univ. Of Cambridge, 1995)

Implementation – for Each ANN (Using Exisiting Systems with New FE)

Build one NN for each of the 60 phone attributes (frame labels obtained from word transcriptions and forced alignment, give 7.6% WER on 5k WSJ)



Window of 9 MFCC frames in order to capture a whole event (about 100 ms)

117 Input nodes

50 Hidden nodes

2 Output nodes

Two output nodes set to complementary values 1 and 0, depending whether the distinctive feature associated to each NN is present or not in the center input frame

Knowledge-Based Features for LVCSR

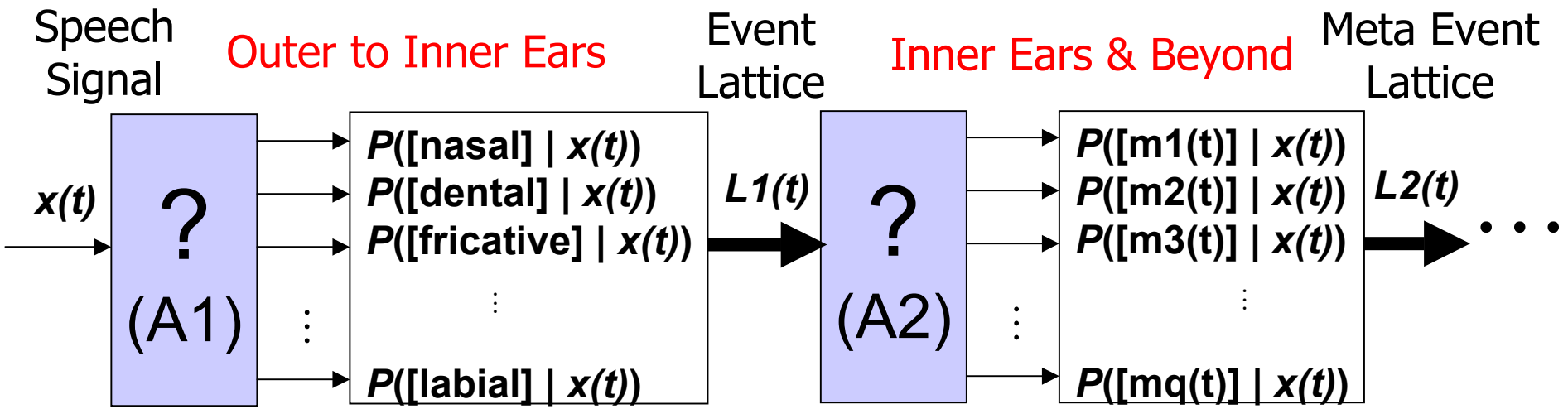
(Integrated with HMM, *Launay, et al. ICASSP02)*

Features/Events	Test 5k	Test 20k
1. Baseline (MFCC, 10msec frame, 39-dim)	4.6	11.8
2. 60 Phonetic Attributes (61-dim)	7.6	16.8
3. 44 Phone Features (45-dim)	5.6	13.0
ROVER Combination without Baseline (2+3)	4.4	11.8
ROVER Combination with Baseline (1+2+3)	3.7	10.6

Word error rates for various feature sets and combinations on WSJ Nov-92, 5k and 20k

- Errors are complementary (key results), but with same MFCC?
- Fuzzy labels and segmentation: main sources for improvement
- Done in 3 months by a visitng MS students using existing tools

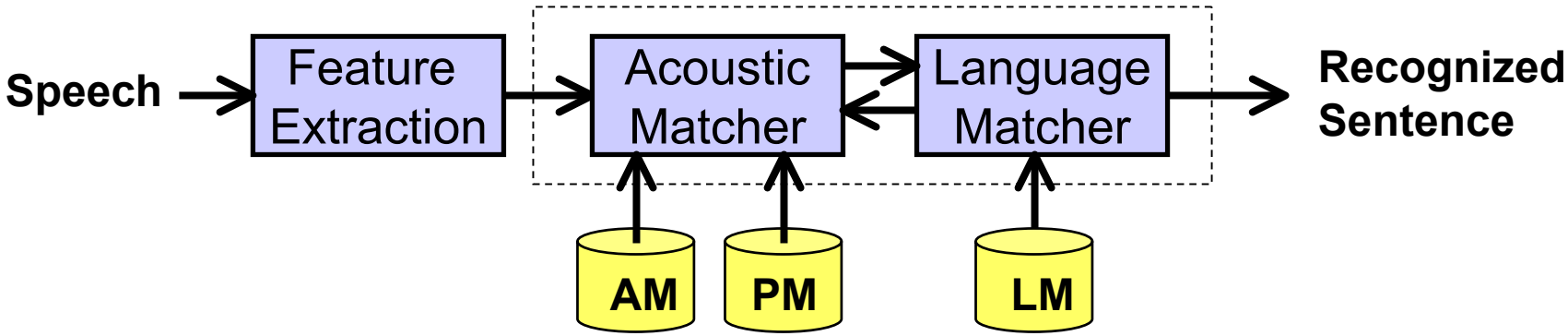
Sequential Detection of Meta Events



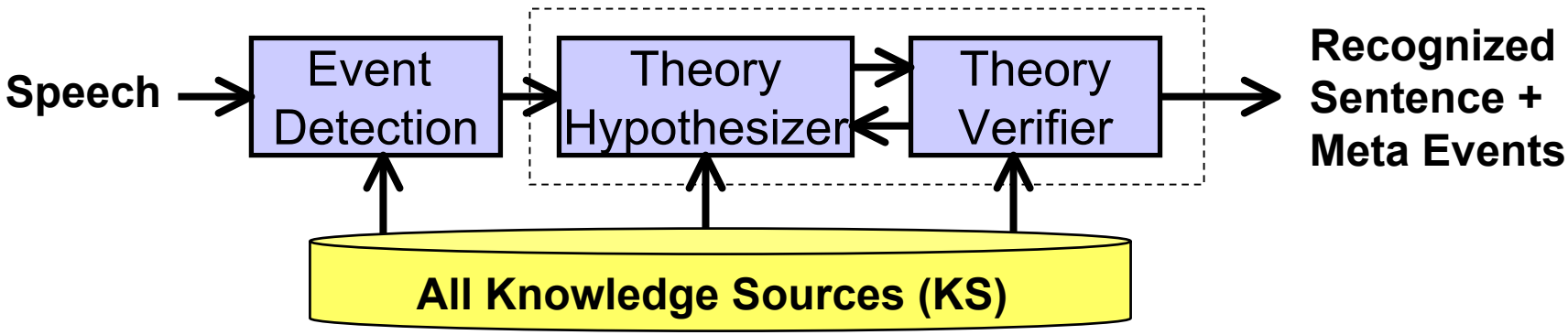
- Each event is modeled by a time series of activity levels, mimicking *neural perception* (CM available at every stage of processing)
- Detection of higher level meta events comeing from coordination of lower level events integrating over space and time (perception?)
- Perceptually inspired signal processing, both analog and digital

An Evolving and Changing Paradigm: From Top-Down Decoding to Bottom-Up Detection

- **Recognition Approach (+ MAP Decoding)**



- **New Detection Approach (+ Lattice Parsing & Verification)**



Summary and Future Work

- **Current Paradigm only achieves incremental progress**
- **New Paradigm with Knowledge Source Integration**
  - better accuracy and meaningful errors for LVCSR
  - new paradigm to address HMM-ASR deficiency
  - new framework utilizing vast knowledge in speech, linguistics and acoustics while advancing statistical modeling paradigm
  - new opportunities for research collaboration
  - faster algorithms combining analog and digital implementations
- **Future Work (plenty of innovative research needed)**
  - open platform for plug-'n'-play to lower ASR entry barriers
  - KS integration via combined event detection and verification
  - more side products with meta info for speech & speaker mining