From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next-Generation ASR

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

Academia Sinica, Taipei, 11/27/04
Outline

• Knowledge-ignorant modeling and ASR past successes
  – Two perspectives: (1) decision theory, (2) channel modeling
  – Advances in speech and language modeling with HMM

• Limitations with state-of-the-art HMM-based systems
  – Robustness, coverage, blackbox, HSR performance gap, etc.

• Knowledge-supplemental modeling
  – Few examples of KS incorporation into state-of-the-art systems

• Knowledge-Rich Modeling: Detection-based ASR
  – Automatic speech attribute transcription (ASAT) paradigm
  – Collaborative research and evaluation platform

• Future work and summary
**ASR: Complete Channel Characterization**

- **P(M)**: message source
- **P(W|M)**: linguistic channel
- **P(S|W,M)**: articulatory channel
- **P(A|S,W,M)**: acoustic channel
- **P(X|A,S,W,M)**: transmission channel

**Sources of Variability**

- *Task specification*
- *Speaker characteristics*
- *Contextual variability*
- *Speaking behavior*

**Speaking environment**

**Transducer variability & distortions**

**Channel variability and distortions**

- message M realized as a word sequence W
- words W realized as a sequence of sound S
- sounds S received by transducer in acoustic ambient as A
- signals A converted from acoustic to electric, transmitted and received as X for processing

- speech recognizer
Knowledge-Ignorant Modeling

Speech Understanding

Noisy Channel

Message \((M, W)\)

Speech \(S\)

Speech \(S\)

Channel Decoding

Estimated \((M, W)\)

Machine Translation

Noisy Channel

Source \(S\)

Target \(T\)

Target \(T\)

Channel Decoding

Estimated \(S\)

Text (Topic) Categorization

Noisy Channel

Topic \(T\)

Document \(D\)

Document \(D\)

Channel Decoding

Estimated \(T\)

Speaker or Language Identification

Noisy Channel

SID (LID) \(K\)

Speech \(S\)

Speech \(S\)

Channel Decoding

Estimated \(K\)
### Other Applications in Pattern Recognition

| Application                      | Input               | Output              | $P(I)$                  | $P(O|I)$               |
|----------------------------------|---------------------|---------------------|-------------------------|------------------------|
| Optical Char. Recognition        | Actual Letters      | Noisy Letters       | Character (Letter) LM   | OCR Error Model        |
| Bioinformatics                   | Actual DNA Sequence | Noisy DNA Sequence  | LM of Nucleotides       | Bio-genetic Model      |
| Machine Translation              | Source Sentence     | Target Sentence     | Source LM               | Translation Model      |
| Text Understanding               | Semantic Concept    | Word Sequence       | Concept LM              | Semantic Model         |
| Part-of-Speech Tagging           | POS Tag Sequence    | Word Sequence       | POS Tag LM              | Tagging Model          |
| Parsing                          | Parse Tree          | Word Sequence       | LM of Derivations       | Parsing Model          |
ASR Capabilities

- Use powerful data-driven modeling tools: HMM, ANN, GM
- Rely little on detailed speech and language knowledge sources
- Give high performance in matched conditions
- Develop and deploy many applications and services
  - but robustness and rigid constraints are two major limiting factors
Shannon’s Channel Modeling Paradigm

- Channel input is hidden (unobserved) while output is observed and used to infer the input (which is often approximated by a structural Markov model)
- Channel modeling with \((I, O)\) pairs in large training sets

\[
\hat{I} = \arg \max_{I \in \Omega} P(I \mid O) = \arg \max_{I \in \Omega} \frac{P(O \mid I)P(I)}{P(O)}
\]
Modeling Input-Output Associations

- Hidden Markov Model (HMM)
- Artificial Neural Network (ANN)
- Classification and Regression Tree (CART)
- Support Vector Machine (SVM) and LVQ
- Kernel-based, mixture of experts, Bayesian network
- Many New Applications
  - Rule induction, statistical parsing, machine translation, etc.
  - Information retrieval, text categorization, call routing, transliteration, pronunciation modeling, etc.
Statistical Decision Paradigm

• Given $P(X,W)$, the joint distribution of the signal $X$ and the pattern $W$; and a loss function, $l(W, d(X))$, of making a decision $d(X)$ when the actual pattern is $W$, then the optimal Bayes decision rule implements:

$$d_0(X) = \arg\min_{d(X)} \sum_W l(W, d(X)) \cdot P(W \mid X)$$

• If $l(W, d(X))$ is a 0-1 loss function, i.e. error count, we have the well-known maximum a posteriori decision rule:

$$d_{01}(X) = \arg\max_W P(X \mid W) \cdot P(W)$$
HMM Estimation – 3 Key Advances

1. **Detailed Modeling** (in many textbooks & packages)
   - More data, context, mixtures, tying, structures, etc.

2. **Adaptive Modeling** (many advances)
   - Adaptation and compensation from extra learning data
   - Coping with new conditions and unexpected situations

3. **Discriminative Modeling** (many advances)
   - From density estimation to decision boundary adjustment
   - Computing models as discriminants, not source distributions
   - Consistent training and recognition/verification objectives

Summary of Modeling Advances

Indirect Parameters

Direct Parameters

Joint Parameters

Some ASR Limitations

- **Unknown joint distributions: optimality?**
- **Robustness and generalization**
  - Mismatch problems in all intermediate channels in ASR
- **Coverage: complete specification not available**
  - High performance in every diverse condition is questionable
- **Open-set versus closed-set ASR**
  - OOV, extraneous speech, “ill-formed utterances”, etc.
- **Blackbox: hard to integrate KS, always re-training**
  - Collaboration opportunities missed, literatures not explored
- **Performance gap between ASR and HSR**
  - Progress slowed down, but HSR ceiling is far away
Robustness: Incomplete Specifications

Signal Space

Training

Feature Space

Testing

Model Space

S → Y → Λ_y

D1(.) → D2(.) → D3(.)

T → X → Λ_x

Feature extraction
Human Speech Recognition (HSR) vs. ASR (Why? Is Ignorance Modeling Enough?)

- Machines Outperform Humans
- Digits
- RM-LM
- NAB-mic
- WSJ
- RM-null
- NAB-omni
- SWBD
- WSJ-22dB
ASR Technology Progress: “S” Curve

Human Performance

ASR Capability

Pattern Matching
(HMM+ANN+Data)
(ignorance modeling)
(good performance)

What’s Next?
Back to Speech Science?
(+ statistical inference)

Speech Science
(before 1970)
(observing specifics)

Knowledge-Supplemental Modeling

• Two-stage “teen-ty” and “e-set” discriminators
  – second-pass detailed classification (e.g. nasal and stop detectors)
• ASR system as a not-so-good “foreign hear”
  – spotting keywords in fluent speech is not hard for HSR and ASR
• Knowledge-based features for LVCSR (WSJ)
• Knowledge scores and lattice re-scoring

KS helps state-of-the-art ASR systems !!
1. Discriminative Feature: Why just MFCC?

- VOT discriminates stops (Ramesh & Niyogi, *ICSLP98*)
- Sound-specific features & detectors (matched filters?)
- Analog and bio-inspired signal processors
  - Beyond frame-synchronous processing?
Two Straightforward Examples:
1. ty-teen discrimination (nasal detectors): 45% less errors
2. e-set recognition (stop detectors): 50% less errors
2. Mimicking “Foreign Ears”
(Key Phrases as Fundamental Attributes)

- ASR with Phrase Detection
  - Phrase Lattice
  - Phrase Model
  - Speech
  - Recognized Results

- Phrase Verification
  - AM & PM

- Sentence Rescoring
  - Task Constraints
  - Parsed Hypotheses

- Lattice Parsing (with Linguistic KS)
  - Reduced Lattice
  - Language Model

Greatly improve semantic accuracy for ill-formed utterances
3. Knowledge-Based Features

- Mimic neural activities with non-linear event detectors
- Simulate posterior probabilities or confidence measures
- Training HMM-based classifiers and verifiers
LVCSR Implementation

Speech Signal ➔ MFCC Extraction ➔ ANN Training ➔ Karhunen Loeve Transform ➔ HMM

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

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

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
Once the WTN composite has been generated, the WTN is searched by majority voting with frequency of occurrences.
### Knowledge-Based Features for LVCSR
(Integrated with HMM, Launay, et al. ICASSP02)

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

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

- Phonetic attributes used in decision tree clustering (Odell ’95)
- Errors are complementary (key consequence) with same MFCC?
A Set of 60 Phone Features (for Clustering)

<table>
<thead>
<tr>
<th>General</th>
<th>Vowels</th>
<th>Consonants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop</td>
<td>Front Vowel</td>
<td>Unvoiced</td>
</tr>
<tr>
<td>Nasal</td>
<td>Central Vowel</td>
<td>Voiced</td>
</tr>
<tr>
<td>Fricative</td>
<td>Back Vowel</td>
<td>Front Consonant</td>
</tr>
<tr>
<td>Liquid</td>
<td>Long</td>
<td>Central Consonant</td>
</tr>
<tr>
<td>Vowel</td>
<td>Short</td>
<td>Back Consonant</td>
</tr>
<tr>
<td>Front</td>
<td>Diphthong</td>
<td>Fortis</td>
</tr>
<tr>
<td>Central</td>
<td>Front Start</td>
<td>Lenis</td>
</tr>
<tr>
<td>Back</td>
<td>Fronting</td>
<td>Neither Fortis or Lenis</td>
</tr>
<tr>
<td>Noise</td>
<td>Medium</td>
<td>Non Coronal</td>
</tr>
<tr>
<td>Silence</td>
<td>Low</td>
<td>Anterior</td>
</tr>
</tbody>
</table>

4. Knowledge Scoring  
(Li, Tsao and Lee, ICASSP05)

Knowledge score: measure a goodness-of-fit between a speech frame and an individual knowledge source

FE\_i : feature extraction module
Y\_i : speech parameter vector
SC\_i : attribute scoring module
KS\_i : knowledge score

FE
SC

FE
SC

FE
SC

FE
SC

FE\_i : feature extraction module
Y\_i : speech parameter vector  
SC\_i : attribute scoring module  
KS\_i : knowledge score

MFCCs
ANN Classifiers

p(A\_i|x(t))

Y\_1(t)

Y\_2(t)
N-Best Rescoring

- Compute scores for places and manners of articulation
- Improve relative continuous phone recognition rate by about 24%
Key Questions

• How to integrate KS in “blackbox” modeling?
  – Learn from multidisciplinary expertise and literatures

• How to lower ASR entry barriers?
  – Encourage individual researchers and small groups to contribute

• Can we solve ASR by combining simpler models?
  – Build focused models to work well in every condition?
  – Collect only right data that will help with new understanding?

• How about partial understanding from spotty clues?
  Do we have a complete KS specification to perform decoding in all cases, including ill-formed utterances?

Next generation collaborative ASR paradigm?
Human-Based Modeling

- Human speech recognition (HSR)
- Acoustic landmarks and invariance: Klatt (ARPA)
- Distinctive features: Fant, Stevens
- Speech production, perception, analysis, etc.
- Learning from reading spectrograms and HSR
  - Explore knowledge hierarchy, from acoustics to pragmatics
  - Detect acoustic and auditory evidences in a speech sound
  - Weigh them and combine them to form cognitive hypotheses
  - Validate them until consistent decisions can be reached

Strong Message: bottom-up KS integration
  - but also taking advantage of data-driven modeling techniques
Missing Links

• Performance gaps between ASR & HSR
• Definition a set of linguistic evidences for ASRU, and a corresponding collection of speech attributes that can be reliably and robustly detected from speech signals
• Quantification and computation of attributes & events
• Formation of high level evidences from low level events
• Decision based on redundant and incomplete knowledge
• Formulation under a rigorous mathematical framework to combine data-driven and knowledge-based approaches

Next generation collaborative ASR paradigm?
Review of Basics

- **Speech** is composed of a sequence of sounds
- **Sounds** (and transitions between them) serve as a symbolic representation of information to be shared between humans (or humans and machines)
- Arrangement of sounds is governed by rules of **language** (constraints on sound sequences, word sequences, etc)--/spl/ exists, /sbk/ doesn’t exist
- **Linguistics** is the study of the rules of language
- **Phonetics** is the study of the sounds of speech

We can exploit *knowledge* about the structure of sounds and language
Human Vocal Apparatus

- **Vocal tract** — dotted lines in figure; begins at the glottis (the vocal cords) and ends at the lips
  - consists of the pharynx (the connection from the esophagus to the mouth) and the mouth itself (the oral cavity)
  - average male vocal tract length is 17.5 cm
  - cross sectional area, determined by positions of the tongue, lips, jaw and velum, varies from zero (complete closure) to 20 sq cm

- **Nasal tract** — begins at the velum and ends at the nostrils

- **Velum** — a trapdoor-like mechanism at the back of the mouth cavity; lowers to couple the nasal tract to the vocal tract to produce the nasal sounds like /m/ (mom), /n/ (night), /ng/ (sing)
Abstractions of Physical Model

excitation → Time-Varying Filter → speech

voiced unvoiced mixed
Speech Sounds

“Should we chase”

/sh/ sound

/ould/ sounds

/we/ sounds

/ch/ sound

/a/ sound

/s/ sound

- Hard to distinguish weak sounds from silence
- Hard to segment with high precision => don’t do it when it can be avoided
Waveform and Spectrogram

SHOULD WE CHASE

- s
- u
- d
- w
- i
- c
- e
- s

FREQUENCY (Hz)

- 5000
- 4000
- 3000
- 2000
- 1000

AMPLITUDE

- 12336
- 6168

TIME (sec)

- 1.0
- 0.8
- 0.6
- 0.4
- 0.2
- 0.0

100 msec
Making Speech “Visible” in 1947

Visible Speech

by Ralph K. Potter
George A. Kopp
Harriet Green Kopp

Center of Signal and Image Processing
Georgia Institute of Technology
Spectrogram Properties

- **Speech Spectrogram** — sound intensity versus time and frequency
- **Wideband spectrogram** - spectral analysis on 15 msec sections of waveform using a broad (125 Hz) bandwidth analysis filter, with new analyzes every 1 msec
  - spectral intensity resolves individual periods of the speech and shows vertical striations during voiced regions
- **Narrowband spectrogram** - spectral analysis on 50 msec sections of waveform using a narrow (40 Hz) bandwidth analysis filter, with new analyzes every 1 msec
  - narrowband spectrogram resolves individual pitch harmonics and shows horizontal striations during voiced regions
Sound Spectrogram: An Example

Every salt breeze comes from the sea.

Frequency (Hz)

Amplitude

Time (sec)
Parametrization of Spectra

- Human vocal tract is essentially a tube of varying cross sectional area, or can be approximated as a concatenation of tubes of varying cross sectional areas.

![Diagram of vocal tract as a concatenation of tubes](image)

- Acoustic theory shows that the transfer function of energy from the excitation source to the output can be described in terms of the natural frequencies or resonances of the tube.

- Resonances known as formants or formant frequencies for speech and they represent the frequencies that pass the most acoustic energy from the source to the output.

- Typically there are 3 significant formants below about 3500 Hz.

- Formants are a highly efficient, compact representation of speech.
Spectrogram and Formants

Key Issue: reliability in estimating formants from spectral data
Acoustic Theory Summary

- Basic *speech processes* — from ideas to speech (production), from speech to ideas (perception)
- Basic *vocal production mechanisms* — vocal tract, nasal tract, velum
- *Source of sound flow* at the glottis; output of sound flow at the lips and nose
- *Speech waveforms and properties* — voiced, unvoiced, silence, pitch
- *Speech spectrograms and properties* — wideband spectrograms, narrowband spectrograms, formants
Linguistic Units of Speech

- Phoneme (abstract, smallest)
- Morpheme
- Syllable
- Word
- Phrase
- Sentence
- Paragraph
- Topics, Articles, Stories
- Others
Traditional Phoneme Classification Chart

Vocal Cords
Vibrating

Noise-Like
Excitation
Distinctive Features (Speech Attributes)

• Classify non-vowel/non-diphthong sounds in terms of distinctive features
  – **Place of articulation**
    • Bilabial (lips)—p,b,m,w
    • Labiodental (between lips and front of teeth)—f,v
    • Dental (teeth)—th,dh
    • Alveolar (front of palate)—t,d,s,z,n,l
    • Palatal (middle of palate)—sh,zh,r
    • Velar (at velum)—k,g,ng
    • Pharyngeal (at end of pharynx)—h
  – **Manner of articulation**
    • Glide—smooth motion—w,l,r
    • Nasal—lowered velum—m,n,ng
    • Stop—constricted vocal tract—p,t,k,b,d,g
    • Fricative—turbulent source—f,th,s,sh,v,dh,z,zh,h
    • Voicing—voiced source—b,d,g,v,dh,z,zh,m,n,ng,w,l,r
    • Mixed source—both voicing and unvoiced—j,ch
    • Whispered—h
The brain recognizes sounds by doing a distinctive feature analysis from the information going to the brain. These features are somewhat insensitive to speaker, noise, background, reverberation => they are robust and reliable.

**FIGURE 17.7** Binary distinctive feature set of Jakobson et al. From [10].

<table>
<thead>
<tr>
<th>Place</th>
<th>p</th>
<th>k</th>
<th>t</th>
<th>b</th>
<th>d</th>
<th>g</th>
<th>f</th>
<th>thin</th>
<th>s</th>
<th>sh</th>
<th>v</th>
<th>the</th>
<th>z</th>
<th>azure</th>
<th>m</th>
<th>n</th>
<th>ng</th>
<th>l</th>
<th>r</th>
<th>w</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>bilabial</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>labiodental</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>dental</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>alveolar</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>palatal</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>velar</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>pharyngeal</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Manner**

| glide       | - | - | - | - | - | - | - | -    | - | - | - | -    | - | -      | - | + | +  | + | + | - |
| nasal       | - | - | - | - | - | - | - | -    | - | - | - | -    | - | +      | + | - | -  | - | - | - |
| stop        | + | + | + | + | + | - | - | -    | - | - | - | -    | - | -      | - | - | -  | - | - | - |
| fricative   | - | - | - | - | - | + | + | +    | + | + | + | +    | + | +      | - | - | -  | - | - | - |
| voicing     | - | - | + | + | + | - | - | -    | + | + | + | +    | + | +      | + | + | +  | + | + | + | + |

*Distinctive English Phoneme Features*

Center of Signal and Image Processing
Georgia Institute of Technology
**Distinctive Features of Speech**

<table>
<thead>
<tr>
<th>Place of articulation</th>
<th>Glide</th>
<th>Nasal</th>
<th>Stop</th>
<th>Manner of articulation</th>
<th>Fricative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilabial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labiodental</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dental</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alveolar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palatal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Back</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Velar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pharyngeal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glottal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Glide</th>
<th>Nasal</th>
<th>Voiced</th>
<th>Unvoiced</th>
<th>Voiced</th>
<th>Unvoiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front</td>
<td></td>
<td></td>
<td>b</td>
<td>p</td>
<td>v</td>
<td>f</td>
</tr>
<tr>
<td>Middle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dental</td>
<td></td>
<td></td>
<td>δ</td>
<td>θ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alveolar</td>
<td></td>
<td></td>
<td>d</td>
<td>t</td>
<td>z</td>
<td>s</td>
</tr>
<tr>
<td>Palatal</td>
<td></td>
<td></td>
<td>r</td>
<td></td>
<td>3</td>
<td>j</td>
</tr>
<tr>
<td>Back</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Velar</td>
<td></td>
<td></td>
<td>w, m</td>
<td>η</td>
<td>g</td>
<td>k</td>
</tr>
<tr>
<td>Pharyngeal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glottal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>h</td>
</tr>
</tbody>
</table>

**FIGURE 17.8** Articulatory classification of consonants. From [15].

Place and manner of articulation completely define the consonant sounds, making speech perception robust to a range of external factors.
Vowel Articulatory Shapes

- Tongue hump position (front, mid, back)
- Tongue hump height (high, mid, low)
  - /IY/, /IH/, /AE/, /EH/ => front => high resonances
  - /AA/, /AH/, /AO/ => mid => energy balance
  - /UH/, /UW/, /OW/ => back => low frequency resonances
The Vowel Triangle

Centroids of common vowels form clear triangular pattern in F1-F2 space

iy-ih-eh-ae-uh
**Diphthongs**

- Gliding speech sound that starts at or near the articulatory position for one vowel and moves to or toward the position for another vowel
  - /AY/ in *buy*
  - /AW/ in *down*
  - /EY/ in *bait*
  - /OY/ in *boy*
  - /OW/ in *boat* (usually classified as vowel, not diphthong)
  - /Y/ in *you* (usually classified as glide)
Unvoiced Fricatives
Nasal Sounds

Hole in spectrum
Voiced and Unvoiced Stop Consonants

- Sounds-/B/, /D/, /G/ (voiced stop consonants) and /P/, /T/ /K/ (unvoiced stop consonants)
  - voiced stops are transient sounds produced by building up pressure behind a total constriction in the oral tract and then suddenly releasing the pressure, resulting in a pop-like sound
    - /B/ constriction at lips
    - /D/ constriction at back of teeth
    - /G/ constriction at velum
  - no sound is radiated from the lips during constriction => sometimes sound is radiated from the throat during constriction (leakage through tract walls) allowing vocal cords to vibrate in spite of total constriction
  - stop sounds strongly influenced by surrounding sounds
  - unvoiced stops have no vocal cord vibration during period of closure => brief period of fraction (due to sudden turbulence of escaping air) and aspiration (steady air flow from the glottis) before voiced excitation begins
Unvoiced Stop Consonants

Stop Gap

\[ \text{uh-\{p,t,k\}-a} \]
Recap

- **Sounds** of phonemes, syllables, words
- **Phonetic transcriptions** of words and sentences — coarticulation across word boundaries
- **Vowels and consonants** — their roles, formants, articulatory shapes, waveforms, spectrograms (manners and places of articulation)
- **Distinctive feature** representations of speech
- **Speech Attributes and Their Transcriptions**: a new way to view and recognize speech
A New Collaborative ASR Paradigm

Paradigm Shift: Automatic Speech Attribute Transcription (ASAT): A new 4-year NSF ITR project at Georgia Tech, Ohio State and Rutgers
5. Shared platform for collaborative system design and objective evaluation
1. Bank of Speech Attribute Detectors

- Speech attributes: “hidden” information embedded in speech to be detected and used as clues for ASRU
  - Linguistic interpretation of fundamental speech sounds
  - Speaker profile: gender, accent, speaking rate, emotion, etc.
  - Speaking environment: interaction between speech & acoustics

- Quantification of speech attributes as “events”
  - Probabilistic description: characterize an event as a time series
  - Computational description: combine events to form other events
  - Confidence description: decide based on evidences
  - Neural activity description
Extraction of Speech Attributes

- ANN-based Attribute Detectors
- Sound-specific parameters & detectors: e.g. VOT
- Biologically-motivated processors and detectors
  - Analog detectors, short-term vs. long-term detectors
  - Why just 10-msec synchronized MFCC?
- Perceptually-motivated processors and detectors
  - Convert speech into neural activity level functions
- Language-dependency vs. language independence
  - Syllable and tone detectors in Mandarin
- Robust detector vs. robust ASR: divide-’n’-conquer
Another Form of Visible Speech?
Language-Independent Detectors?

Stop

Nasal

Vowel

.....
2. Event Merger

• Merge multiple time series into another time series
  – Maintaining the same detector output characteristics
  – Enhancing evidences and suppressing noise

• Combine temporal events
  – An example: combining phones into words (word detectors)

• Combine spatial events
  – An example: voiced and stop features into voiced stop event

• Extreme: build a 20K-word recognizer with 20K keyword detectors
Event Merging Under Speech KS Hierarchy (Related to Auditory Perception?)

- Each event is modeled by a time series of activity levels, mimicking *neural perception* (confidence measure available at every stage)
- Detection of higher level meta events comes from coordination of lower level events integrating over space and time (perception?)
3. Evidence Verifier

- Provide confidence measures to events and evidences
  - Utterance verification algorithms can be used
- Output recognized evidences (words and others)
  - Hypothesis testing is needed at every stage
- Prune event and evidence lattices
  - Threshold for decision based on confidence measures
- Design minimum verification error (MVE) verifiers
- Need many new theories
- Others?
From Classification To Verification

- Probability $P(T(X) \mid H_0)$
- Probability $P(T(X) \mid H_1)$

Regions:
- Region I: Target
- Region II: Impostor

Threshold
Research Issues in Pattern Verification

• **Unknown H0 and H1 distributions: open set**
  – H0: How to compute $P(X|H0)$?
  – H1: How to approximate $P(X|H1)$?

• **Optimal tests are not available**

• **Robustness and Generalization Issues**

• **Generalized Likelihood Ratio (GLR) Test**
  – Separation between target and competing models
  – Data-driven verification test design
  – discriminative modeling (e.g. MVE)
Verification of Composite Events
(Re-enforcing target, suppressing impostors)

\[ P(\text{Voiced} \mid X_{t_1}^{t_2}) + P(\text{Stop} \mid X_{t_1}^{t_2}) = P(\text{Voiced Stop} \mid X_{t_1}^{t_2}) \]
Verifying Sequentially Combined Events

(/w+/Λ/+/n/ = “one”)
Comparison of Pattern Verifiers

Receiver Operating Characteristic (ROC) Curves for overall comparison!!
4. **Knowledge: Definition & Evaluation**

- Explore large body of speech science literature
  - Focus on quantification, computation, robustness, evaluation
- Define training, evaluation and testing databases
- Develop Objective Evaluation Methodology
  - Defining detectors, mergers, verifiers, recognizers
  - Defining/collecting evaluation data for all
5. **System Prototypes & Common Platform**

- **Continuous Phone Recognition: TIMIT**
- **Continuous Speech Recognition**
  - Connected digit recognition
  - Wall Street Journal
  - Switchboard?
- **Establishment of a collaborative platform**
  - Implement divide-’n’-conquer strategy
  - Build plug-’n’-play modular des’
  - Develop a user community
Objective Evaluation Methodology

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

   - Event-Specific Evaluation Data
   - Library of Detectors
   - New Event Detector
   - Comparator
   - Detection Results
   - Event-Specific Diagnostic Results

2. Diagnostic Evaluation of Speech Recognizers:

   - Event-Intensive ASR Test Data
   - ASR with New Detector
   - ASR with Old Detector
   - Comparator
   - ASR Results
   - Event-Specific Diagnostic Results
An Evolving and Changing Paradigm: From Top-Down Decoding to Bottom-Up Detection

- Convention Recognition Approach (+ MAP Decoding)

Speech → Feature Extraction → Acoustic Matcher → Language Matcher → Recognized Sentence

<table>
<thead>
<tr>
<th>AM</th>
<th>PM</th>
<th>LM</th>
</tr>
</thead>
</table>

- New Detection Approach (+ Lattice Parsing)

Speech → Event Detection → Theory Hypothesizer → Theory Verifier → Recognized Sentence + Meta Events

All Knowledge Sources (KS)
Summary and Future Work

- **Knowledge-ignorant** modeling approach is mathematically well-formulated and carries us a long way in the past
- **Knowledge-supplemental** modeling enhances state-of-the-art ASR systems by exploring knowledge sources
- NGASR: **Knowledge-rich** modeling combines the best of data-driven and knowledge-based approaches
  - Paradigm shift from top down decoding to bottom-up detection
- ASAT paradigm solves ASR with a divide-’n’-conquer strategy and a plug-’n’-play sharable platform
  
  A collaborative ASR community of the 21st Century