An Overview of Automatic Speech Recognition

Slides created by Matt Ferrante
Some slides used from MIT

Sources

Main Algorithm Analyzed:

- Willie Walker , Paul Lamere , Philip Kwok , Bhiksha Raj , Rita Singh , Evandro Gouvea , Peter Wolf , Joe Woelfel, Sphinx-4: a flexible open source framework for speech recognition, Sun Microsystems, Inc., Mountain View, CA, 2004

Slides Used:

- Some created by Matt Ferrante
- Slides From MIT CS 6-345 2003:
- http://ocw.mit.edu/OcwWeb/Electrical-Engineering-and-Computer-Science/6-345Automatic-Speech-RecognitionSpring2003/LectureNotes/index.htm

Outline

- Problem
- Why?
- History
- Applications
- Background
- Challenges / Difficulties
- Metrics
- Algorithm
- Some other Methods
- Future Work
- Conclusions
- References

Problem

- Matching sound to templates for commands is easy enough. Can we dictate to a system, having it record the words that are being spoken?
- Is speech input reliable enough to replace a keyboard?
- Humans can focus on one person talking in a crowded room, how can a computer do this?
- How can we distinguish between different speakers?
- How can we distinguish between ambient noise and someone speaking?
- How can we derive meaning from what was said?
- Can we get what the user meant to say?
  - Some users have elements about their speech that make it difficult to record.

Why?

- The internet is not just text and images anymore, with HTML5, video and audio is supported extremely easily.
- With increasing multimedia on the internet, automatic transcription of sound to text would be very helpful for search
- It would also be useful for watching online videos for the deaf.
- We can typically speak more quickly and efficiently than we can type.
- We communicate with people via speech, why should it be any different for dealing with technology?
- To give us a greater understanding of how the sounds we make build words.
- Sci-fi does it.
- It is pretty cool.

History

- First example of speech recognition was in 1952
  - Could recognize spoken digits
- Was presented as a replacement for keyboard input
  - Failed because is not reliable or accurate enough
  - Only successful once presented as a supplement to keyboard input.
- Error rates started very high, now much more reasonable. Less than 10% in the majority of cases for English.

MIT

ASR Trends*: Then and Now

	before mid 70's	mid 70's - mid 80's	after mid 80's
Recognition Units:	whole-word and sub-word units	sub-word units	sub-word units
Modeling Approaches:	heuristic and ad hoc	template matching	mathematical and formal
	rule-based and declarative	deterministic and data-driven	probabilistic and data-driven
Knowledge Representation:	heterogeneous and complex	homogeneous and simple	homogeneous and simple
Knowledge Acquisition:	intense knowledge engineering	embedded in simple structure	automatic learning

* There are, of course, many exceptions.

Applications

- Translate video and audio into text for Web Search.
- Real - Time or Recorded Functions
  - Translation
  - Captioning
- Automatic Telephone Call Processing
- Augmented Reality
  - Device recording real world conversations you have with other people
- Enhancing User Interfaces
  - Command - Based
  - Dictation
- Accessibility
  - People who can't type due to injury
  - People who can't see keyboard (blind)
  - Children who can't type yet
- Hand-free control in Cars

Background

Speech

- Biological Factors
  - The way our mouths move to produce certain sounds effect the features of the sound itself.
  - The structure of the mouth produces multiple waves in certain patterns.
  - When we manipulate our mouths in the way to make a 't', we push out more air at once, making a higher frequency sound.
- **Phonology**
  - How we use sound to convey meaning in a language
  - In english it states characteristics of sounds like vowels and consonants.

Background

Frequency of Sounds

- Different vowels have different pitches, they are similar to musical notes
  - 'i' being the highest
  - 'u' being the lowest
- Consonant phonemes have more waves oscillating of different parts of the mouth.

Timing

- There is a lot of information in timing.
- Breaks between words have a break in speaking in most cases.
- Vowels last longer than consonants.

Background

Phoneme

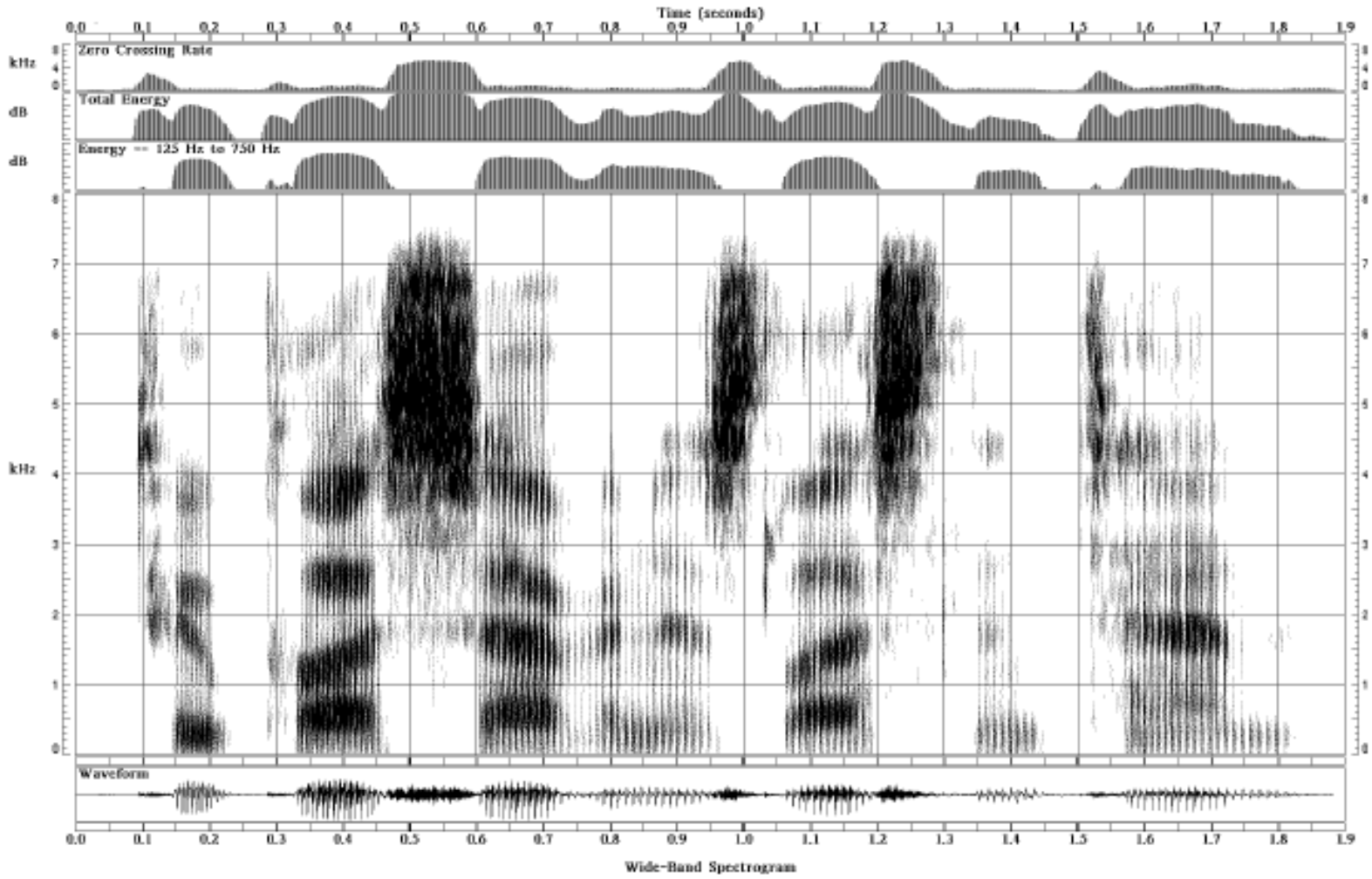
- The smallest segmental unit of sound in a language
- Each Phoneme has features in the sound that differs it from another Phoneme
- Combine to represent words and sentences
- English has about 40 phonemes

Consonant phonemes of English

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar[2]	Palatal	Velar	Glottal
Nasal[1]	m			n			ŋ	
Plosive	p b			t d			k g	
Affricate					tʃ dʒ			
Fricative		f v	θ ð	s z	ʃ ʒ		(x)[3]	h
Approximant				ɹ[1, 2, 5]		j	w[4]	
Lateral				l[1, 6]				

Wikipedia

A Wideband Spectrogram

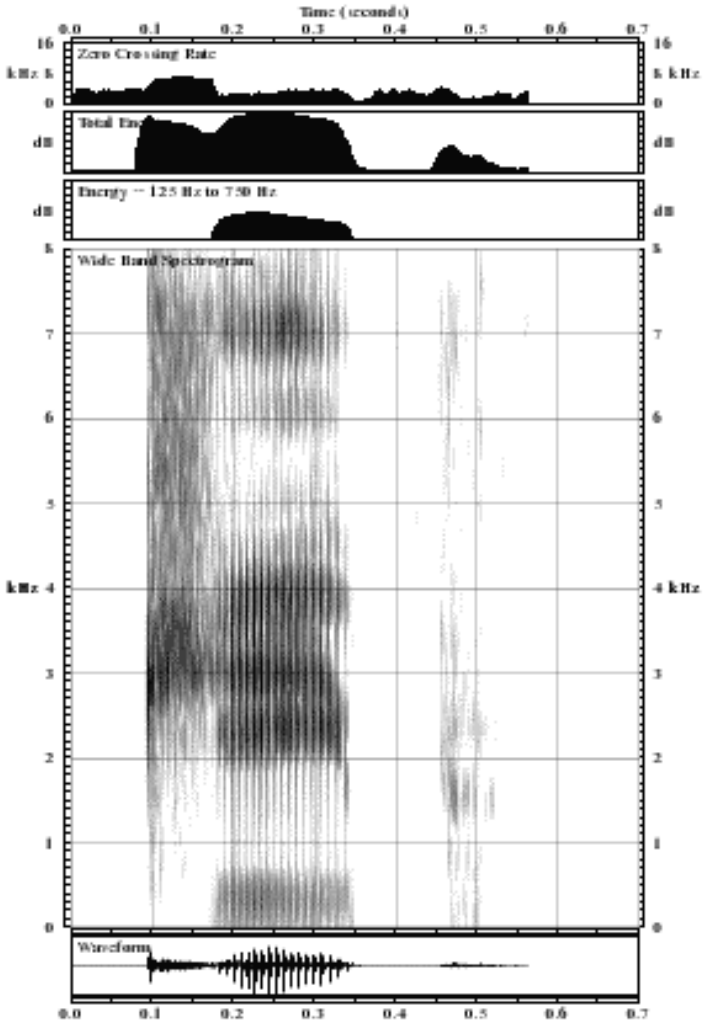


Two plus seven is less than ten

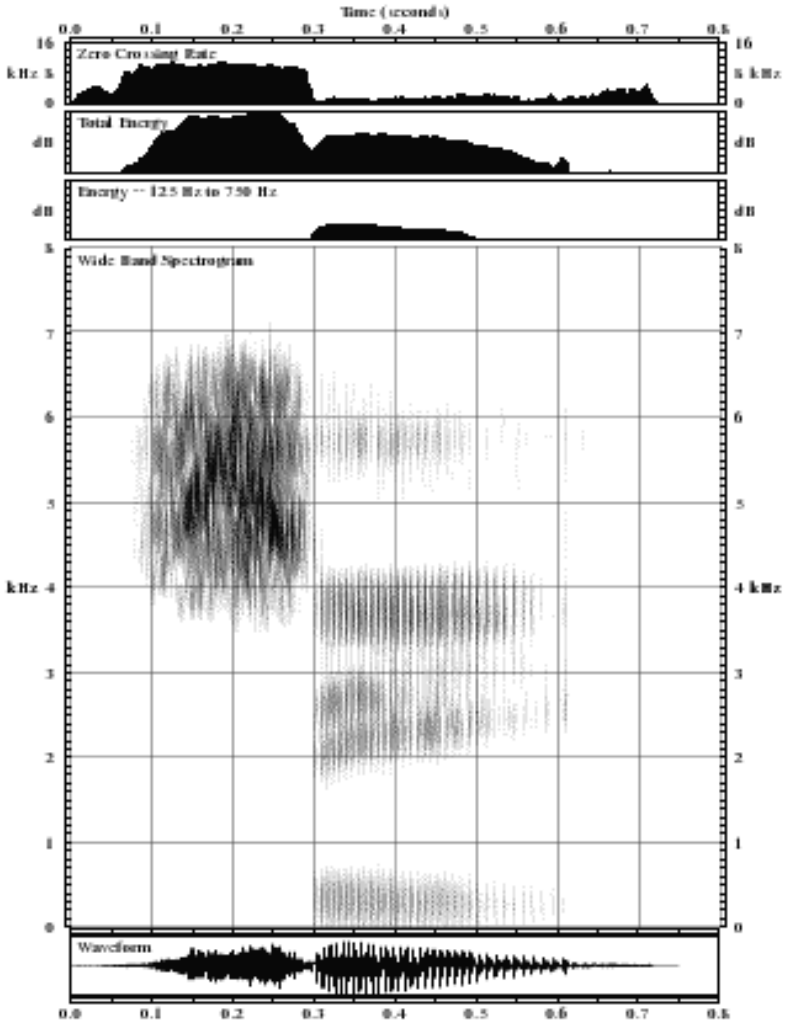
Phonemes in American English

PHONEME	EXAMPLE	PHONEME	EXAMPLE	PHONEME	EXAMPLE
/iʸ/	beat	/s/	see	/w/	wet
/ɪ/	bit	/š/	she	/r/	red
/eʸ/	bait	/f/	fee	/l/	let
/ɛ/	bet	/θ/	thief	/y/	yet
/æ/	bat	/z/	z	/m/	meet
/ɑ/	Bob	/ž/	Gigi	/n/	neat
/ɔ/	bought	/v/	v	/ŋ/	sing
/ʌ/	but	/ð/	thee	/č/	church
/oʷ/	boat	/p/	pea	/ǰ/	judge
/ʊ/	book	/t/	tea	/h/	heat
/uʷ/	boot	/k/	key		
/ɝ/	Burt	/b/	bee		
/ɑʸ/	bite	/d/	Dee		
/ɔʸ/	Boyd	/g/	geese		
/ɑʷ/	bout				
/ə/	about				

Example of Consonant Spectrograms



/kiʸp/



/siʸ/

Challenges / Difficulties

Transcription
- How can we translate from frequencies to a representation of a phoneme?
- What information is kept from the recording, what is discarded?

Correctness
- Was the translation correct?
- How sure are we that we were right?
- Does the sentence created make sense?

Learning
- How can this system learn from its mistakes?

Imperfect Speech
- Stutters
- Saying 'um'

Metrics of Voice Recognition

Evaluating Voice Recognition Algorithms:

- Performance accuracy
  - Correct Words / Total Words
- Word error rate - For sentence context based algorithms
  - Wrong Words / Total Words
- Single word error rate - Raw words from Phonemes
  - Wrong Words / Total Words
- Command success rate
  - For systems that take commands
  - Successful Commends / Commands Issued
- Speed
  - Words / Minute
  - Levels of accuracy for different speeds in WPM

Algorithm

This example is from the Sphinx-4.

- Open source
- Developed at Carnegie Mellon University
- Built for modularity and with research in mind
- Reference 1 on Reference slide

Starts With

- Speech to Feature Engine
- Linguist
  - Acoustic Model
  - Dictionary
    - Words broken into the phoneme sounds they are typically made of.
  - Language Model

Algorithm

General Idea

Training

- Use a training set of many speeches with the texts.
- Extract phonemes from the speeches.
- Build statistical knowledge.
  - Phonemes
  - Words

Recognition

- User speaks
- System extracts features from the speech.
- Those features statistically match up with with a phoneme.
- Use the word statistics to go from phoneme ordering to words.

Algorithm

Speech to Features

- Based on all the features of a sound wave
  - Frequency
  - Pitch
  - Amplitude
  - Time information
- Mathematically give values to the features observed

Algorithm

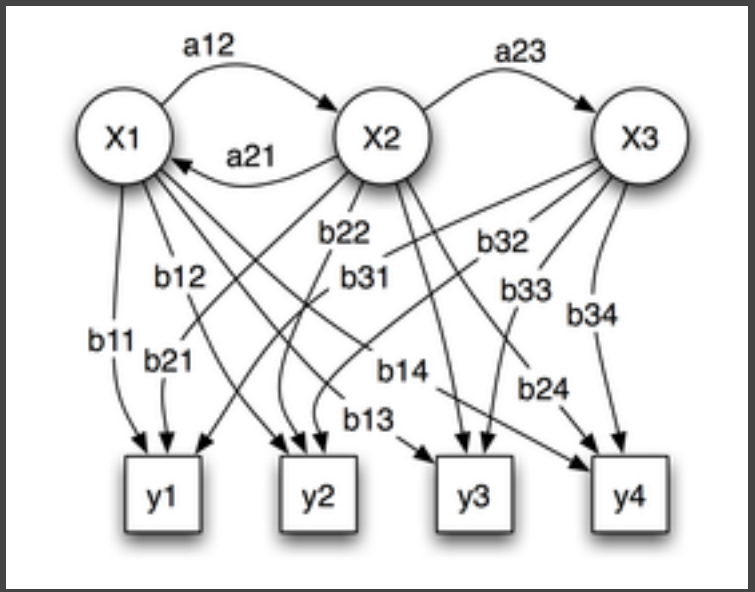
Acoustic Model

- The Acoustic Model is the statistical mapping from the units of speech to all the features of speech.
- Used for Speech Sound to Phoneme
- Used for Phoneme to Word
- Statistical Models
  - Naive Bayes
  - Hidden Markov Models
- It is given information about the language Phonology.
- It can learn from a training set

Algorithm

Hidden Markov Models

- The first state is not known
- The probability of the state X being unit S of speech, based on the features Y that it most likely exhibits, given the values it has for those features and the statistical values.
- Also has probability for its surrounding neighbors.
  - The probability that X1 will be followed by X2
- Some implementations will do probability of other neighbors as well



Wikipedia

X = states
y = possible observations
a = state transition probabilities
b = output probabilities

Algorithm

Language Model

The Language Model

- Provides word-level structure for a language
- Formal Grammar Rules
- Graph Models
  - Word nodes
  - Probability edges from node N to M
  - Probability weight that M will be after N
- N-Gram Models
  - Probability of word is based on the last N-1 terms

Algorithm

Linguist

- The Linguist
  - Acoustic Model
  - Language Model
  - Dictionary
- These elements combined, are the knowledge base for the system
- The linguist knows the language like someone who is fluent would.
- Can make mistakes just like humans do.
  - "What did you say?"
  - 'Lettuce' vs 'Let us'

Algorithm

Linguist Training

- Needs Documents
  - Many documents
  - Wide range of document subjects
    - Covers more words
    - Has more examples
  - For more domain specific set of words, only use related subjects.
    - Good for training a classroom voice recognizer.
  - Many words
    - Large Variety for large word domain.
- Have different people read same documents
  - Get more statistics for a certain word
  - Makes the certainty of a certain word higher during recognition.

Algorithm

Linguist Training

- Needs People
  - One Person
    - Good for your his or her own computer
    - Very accurate for that person
  - Many people
    - Will have much more of a range with different voices
    - Different kinds of people
      - Gender
      - Accents
      - Dialects
- The Linguist is trained and has the knowledge of speech sound to word relations.

Algorithm

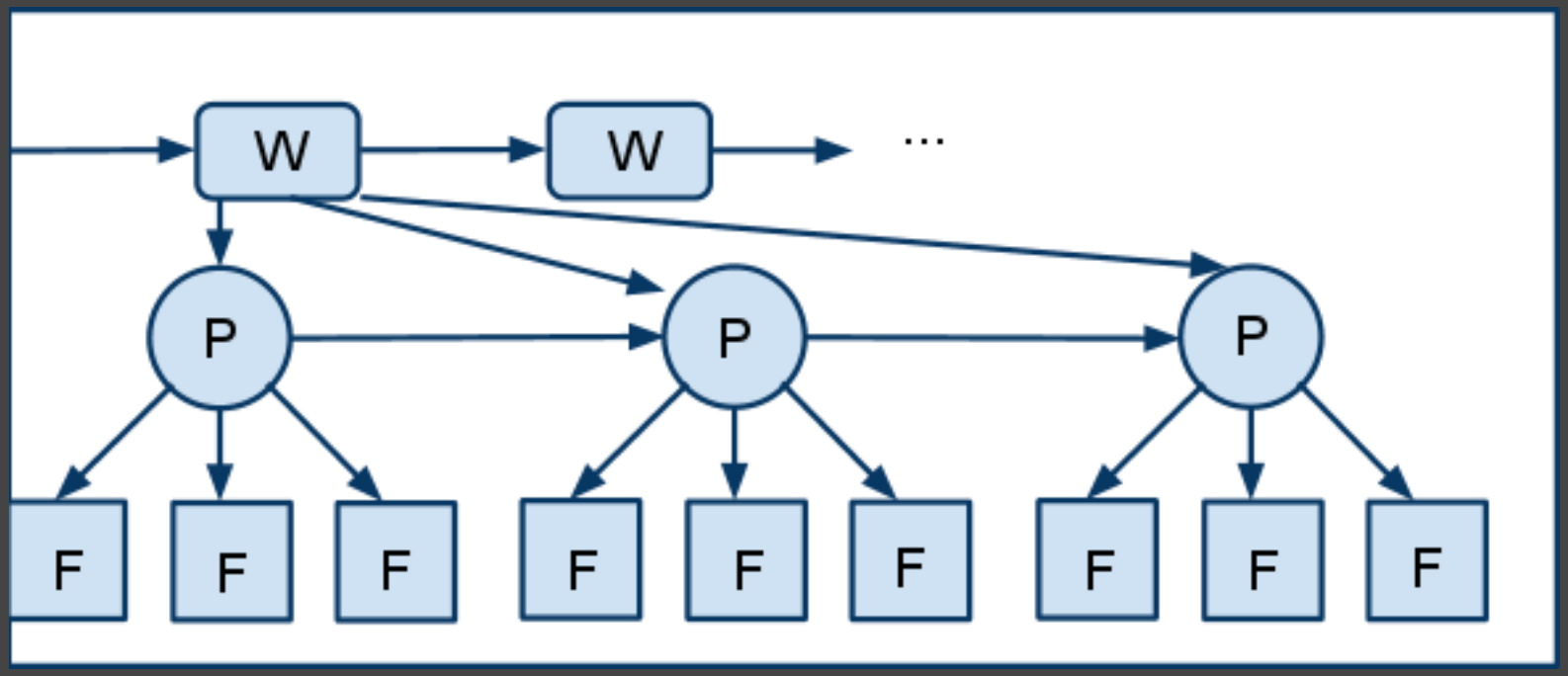
SearchGraph

- The Linguist generates a SearchGraph combining
  - Acoustic Model
  - Dictionary
  - Language Model
- Search Structure contains nodes
  - Low-level states can be scored against the speech features and related probabilistically
  - High-level states represent words and phonemes
- And edges between these nodes
  - Represent possible state transition
  - Have probability value for likeliness transition
- Paper does not go into detail about how exactly the search structure is built.
- Directed graph created by combining the HMMs

Algorithm

SearchGraph

W = unknown word state
P = unknown phoneme state
F = feature



- Probability of a phoneme is based on the previous phonemes and the probability that a feature observed indicates that P is that phoneme
- Probability that W is a certain word is based on on probability that each P is a certain Phoneme

Algorithm

Decoding

On the human level:

- Someone who is fluent in english understands how the sounds work to convey meaning
- A person can understand what is being said by relating the order of sounds to words through the linguistic model that they have built.
- Then derive meaning from those words.

Decoding at a computer level

- We have a set of features over a time interval for spoken sound
- We have a SearchGraph
- We can search this structure for the most probabilistic resulting string based on the features of the speech.

Algorithm

Decoding

- Searching the SearchGraph gives us a set of texts that could represent the speech.
- Each text has a probability associated with it, for how probable it was based on the features found.

The probability of observing a sequence

$$Y = y(0), y(1), \ldots, y(L-1)$$

of length L is given by

$$P(Y) = \sum_{X} P(Y \mid X) P(X),$$

where the sum runs over all possible hidden-node sequences

$$X = x(0), x(1), \ldots, x(L-1).$$

Wikipedia

Other Methods Used

Candidate List / Alternative Hypothesis

- Instead of calculating probability of text, given speech.
- Keep information for post-processing
- Calculate probability of a certain word in a certain spot, given speech.
- Use most probable word
- Keep list of alternative hypothesis for post-processing
- Use knowledge of language as well as heuristics to correct errors after statistically setting up the text.

Error Detection / Prevention

- Using a domain specific linguist, more accurate results are found when talking about that domain.

Other Methods Used

Using Video for Auditory Confirmation

- Better for recognizing when it is a person speaking instead of ambient noise.
- Better for distinguishing between people
- Sensor Fusion increases information by coordinating senses.

Building a robust system for noisy atmospheres

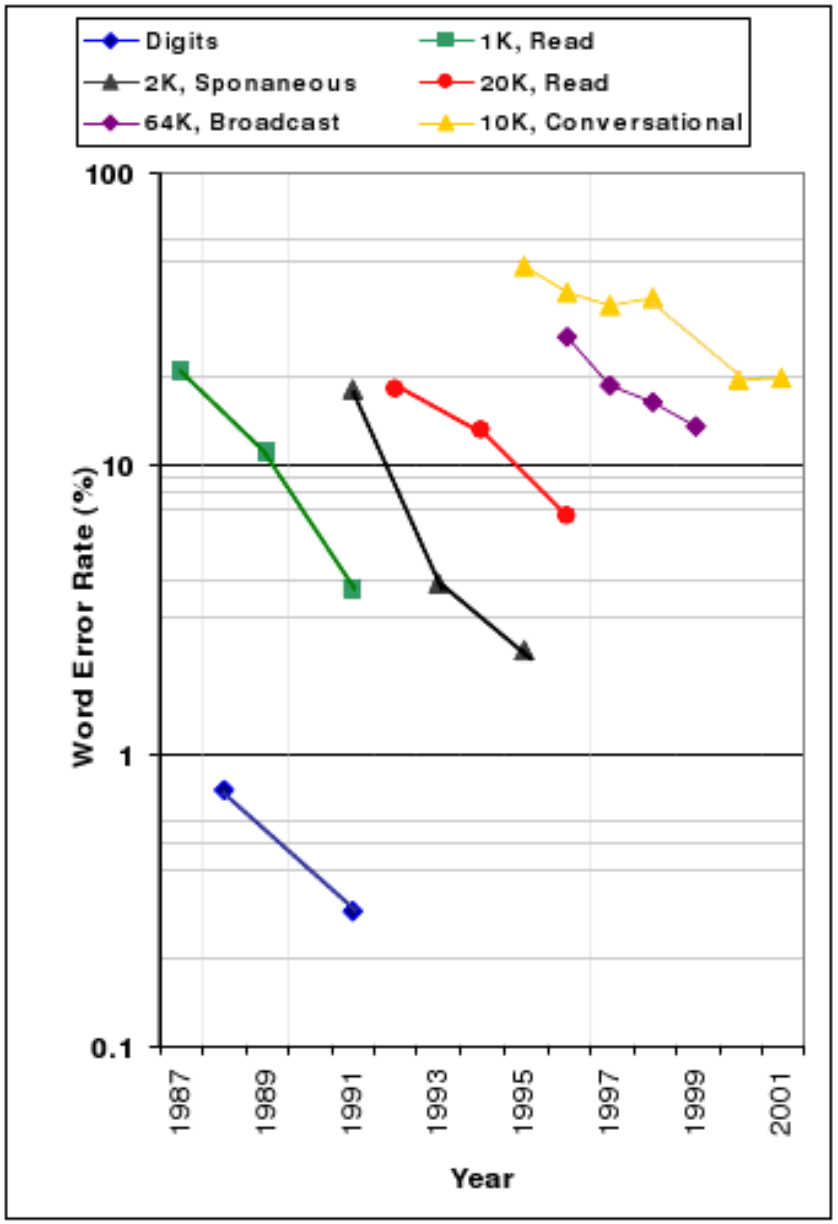
- Filtering and Spectral Subtraction of reoccurring static noise.
- Often the system itself is loud (Cars). Use the information about what noises the car is making, ignore it.

Analyzing the Algorithm

- Good for dictation.
- Non-deterministic, like human speech recognition.
- SearchGraph can be very inefficient because it has to check many combinations
- Does not go so far as to try to extract meaning from the speech
- Does not make an attempt of distinguishing between people.
- Does not keep information about phonemes after conversion to word.

Examples of ASR Performance

- Speaker-independent, continuous-speech ASR now possible
- Digit recognition over the telephone with word error rate of 0.3%
- Error rate cut in half every two years for moderate vocabulary tasks
- Error for spontaneous speech more than twice that of read speech
- Conversational speech, involving multiple speakers and poor acoustic environment, remains a challenge
- Tens of hours of training data to port to a different domain
- Statistical modeling using automatic training achieves significant advances



Future Work

- Languages like Mandarin do not have as high accuracy rates.
- Better error detection.
- Better error fixing after detection.
- Making system more robust.
- What has to be done to make the system to detect the noise that it is making and ignore it?
- How can we extract meaning from what is said?

Conclusions

- Voice recognition is a very important task for indexing videos.
  - Speech -> Text -> Index
  - Search
  - Videos are growing in importance on the internet with YouTube and average people are documenting things. Need to search.
- We have made a lot of progress in the last 50 years, but here is still a lot more to do in the field of Voice Recognition.
- Voice Recognition can make peoples lives easier by proving them a direct line of communication with a machine.

References

1. Willie Walker , Paul Lamere etc... *Sphinx-4: a flexible open source framework for speech recognition* , Sun Microsystems, Inc., Mountain View, CA, 2004
2. MIT Open CourseWave - Lecture Notes / Slides for 6.345 in 2003. http://ocw.mit.edu/
3. Wikipedia: Speech Recognition, Acoustic Modeling, English Phonology
4. Huang and Alleva, *An Overview of the SPHINX-II Speech Recognition System.* HLT:'93 Proceedings of the workshop on Human Language Technology. 1993
5. Trent W. Lewis , David M. W. Powers, *Sensor fusion weighting measures in Audio-Visual Speech Recognition* , Proceedings of the 27th Australasian conference on Computer science, p.305-314, January 01, 2004, Dunedin, New Zealand
6. John-Mark Bell, *Enhancing accessibility through correction of speech recognition errors,* SIGACCESS NEWSLETTER, ISSUE 89, SEPT 2007