Automatic Speech Recognition Tutorial

Professor Don Colton – BYU Hawaii

June 17, 2003

1 Overview

Automatic Speech Recognition (ASR) is a computerized process. It receives as its input a speech recording. It produces as its output a transcription. In the computing world, ASR is an unsolved problem; nobody has yet demonstrated the ability to do it as well as humans can. This tutorial looks at the automatic speech recognition process, from intention to response, with varying degrees of detail.

This tutorial is presented as an overview and as a study guide for an undergraduate elective course in Automatic Speech Recognition, CS 441, taught at Brigham Young University Hawaii by Don Colton. The document is in an early stage of development and includes many rough-hewn statements that deserve and anticipate more accuracy and supporting citations in future drafts. In particular, it contains statements that I have heard and/or believe to be true, but which I have **not yet verified.** You have been warned.

2 Intention

Speech starts with the intention to communicate. There are many man-made sounds that may or may not involve any intention to communicate: a sigh, a sneeze, a burp, flatulance. We will restrict our attention to sounds that are intentional. The goal of such sounds is typically to cause understanding and/or response in a listener.

There is a related laziness issue. My conjecture is that most communication is only clear enough that the producer has a reasonable expectation that the hearer as understood it. For example take the utterance "djuwanna eat?" This could be perceived as "did you want to eat" or "do you want to eat" even though the did/do is not carefully articulated, and the t in "want to" is gone. Still, there is enough information present that the message gets through.

Also because of the expectation of intentionality of speech, hearers can apply common sense to resolve ambiguity by looking for meaning. If the utterance cannot be physically discriminated between "professors least driving" and "professedly striving" common sense may dictate the latter as more likely.

3 Articulation and Resonance

We also focus our attention on vocalizations, by which we mean sounds that are made using the mouth, nose, and throat. We ignore sounds such as clapping, finger snaps, and knuckle or neck cracking.

3.1 Articulators

Articulation is the act or process of dividing something into separate articles. In this case, the thing divided is sound and the separate articles are phonemes. We create different phonemes by intentional movement of our lips, tongue, jaw, vocal cords, and lungs.

Lung movement creates an air pressure differential that creates air flow through the other articulators. The air flow may be inward (pulmonic ingression) or outward (pulmonic egression).

3.2 Vocalization and Whisper

As the air flows past the vocal cords, tension applied to the vocal cords causes the creation of a tone or pitch. When this tone is present, the speech is said to be voiced. When his tone is absent, the speech is whispered.

All vowels are voiced. Some consonants are voiced (b, d, g, v, z, zh, j) while others are unvoiced (p, t, k, f, s, sh, ch). In whispered speech this difference cannot be detected, and yet humans can generally understand whispered speech. This shows that substantial redundancy is present in speech because the whole signal is not required in order to determine the meaning.

3.3 Chambers

The positioning of tongue, jaw, and lips creates resonant chambers of various sizes within the head. These chambers contribute additional tonality to speech. As you slowly say the word "I," you will find your jaw is open wide at the beginning and nearly closed at the end. The mouth chamber varies in size as the word is uttered.

Resonance can be illustrated by thinking about bottles partly filled with water. If one blows softly across the mouth of the bottle, a tone is produced. The blowing sound itself is similar to the letter f. It consists of a

broad spectrum of sound similar to white noise. Some of the frequencies present will resonate with the chamber in the bottle. These will reinforce each other while the non-resonant frequencies cancel each other. The result is a perceptable tone at a stable frequency. By removing water from the bottle the resonant chamber is made larger. This results in a lower-pitched tone and a longer wavelength for the sound. By adding water the wavelength (distance from the mouth of the bottle to the top of the water) is made shorter, resulting in a higher-pitched tone. A similar principle is used in many musical instruments to vary the pitch of the sound that is created.

4 Hearing

Human hearing consists of several steps. Sound is channeled into the head through the outer ear and the ear canal. The sound is focused on the ear drum (timpanic membrane). It is then transmitted and amplified by three small bones onto the side of the cochlea. The cochlea is a fluid-filled coil of varying diameter. It is filled with hairs (cilia) that sense the vibrations in the fluid.

4.1 Cochlea

Low-pitched sounds cause vibration at the wide end of the cochlea. High-pitched sounds cause vibrarion at the narrow end of the cochlea. Just as with the resonant bottle mentioned above, high-pitched sounds do not cause vibration at the wide end because they cancel each other out. The progession from wide to narrow allows the cochlea to discriminate between a large range of sounds, each of which will find a resonance somewhere.

4.2 Nerve Signals

The cilia within the cochlea are connected to nerves. As the cilia vibrate, the nerves fire signals into the brain. The brain combines the signals (sensory fusion) from various nerves to develop an overall sensation of the sound that is being received.

4.3 Frequency Range

The frequency range of sounds perceived by humans is something between 30 Hz (hertz, cycles per second) and 20,000 Hz. Above 20K we hear nothing. Dog whistles operate at that range, where humans do not hear but dogs still do. Below 30 Hz we distinguish separate beats rather than a fused (combined) tone. The limits vary by person. As we grow older, our top frequency tends to get lower.

4.4 Frequency Response

Human hearing is not equally sensitive at all frequencies. At very high frequencies, a sound must be louder to be perceived. The same is true at very low frequencies. The optimal frequence is about 1000 Hz. A sound at 1000 Hz is more easily heard than another sound with the same amount of energy at any other frequency.

5 Microphone

In a speech recognition system, a microphone substitutes for the ear drum. The pressure waves present in the air are transmitted into an analog electrical signal, perhaps by varying the density of carbon dust in a box, or by some other means.

5.1 Frequency Response

The best microphones have a uniform frequency response, meaning that they "hear" sounds equally well from a broad range of frequencies. For speech recognition, the microphone should have a good frequency response throughout the normal human hearing range.

5.2 Noise Cancellation

One feature of high-quality microphones is the ability to cancel noise. This is often done by using two microphones, with one directed toward the sound of interest and the other oriented in the opposite direction. Ambient (nearby) noise should be heard about the same in both microphones. The sound of interest should be louder in the primary microphone. By subtracting the two signals, noice is cancelled out, leaving only the signal with the greater differential. This improves the signal-to-noise ratio for the microphone.

5.3 Spectral Subtraction

Another method of noise cancellation is to look for stable noise frequencies present in the environment. These can come from fluorescent light ballasts, computer fans, the 60 Hz hum of the electrical system, or the slow whistle of air conditioning vents. Because these sounds are stable, we can improve the signal-to-noise ratio for speech by deleting all signals at those stable noise frequencies. This is not as good as noise cancellation based on two microphones, because it loses true signal as well as noise. But it is still an improvement.

6 Digitization

The analog signal from the microphone can be recorded directly to magentic tape using a traditional tape

recorder. The signal can then be reproduced with excellent fidelity (truth) by replaying the tape through an amplifier and speakers.

We are more interested in digital processing of the signal. For this we must convert the incoming voltages into numbers. This is done by sampling the signal many times per second. We record the signal level at each of those times.

6.1 Sample Rate

Sampling is familiar to anyone who has watched a movie or a television program. On TV, the content is painted to the screen at a rate of about 30 frames per second (NTSC). We can notice this when we watch a wheel with spokes as it spins. The faster it goes, the faster we see it go, until a certain point where the blur of the spokes actually appears to slow down and reverse direction. At the right speed, a wheel can appear to be standing still, even though the carriage on which it is mounted is clearly moving forward rapidly.

As the wheel speeds up, we see it speed up. When the wheel reaches a speed where the spoke moves half way to the next position in 1/30 of a second, we cannot tell whether the wheel is moving forward or backward. Beyond that speed, the wheel begins to rotate backward (apparently) until it slows to a stop. A wheel whose spokes moved exactly one position in 1/30 of a second would appear to not be rotating. A wheel whose spokes moved slightly less than one position in 1/30 of a second would appear to be rotating backward.

The interesting thing is that in general human perception is not upset by these 30 frames per second. Instead, we see smooth motion. But this is perhaps less surprising when we realize that the human neural system provides bursts of data anyway. We do not seem to be working with an analog signal.

6.2 Nyquist's Theorem

There is a theorem called Nyquist's Theorem. Basically it states that to correctly resolve a signal at some frequency, you have to sample it at more than twice that frequency.

Applied to wheels and spokes, with a 30 Hz sampling rate, the fastest wheel you can watch is moving one spoke position 15 times per second. Actually, at exactly 15, you cannot tell whether the wheel is going forward or backward.

Applied to speech, it means a signal at 1000 Hz must be sampled at more than 2000 Hz or it cannot be resolved into its correct values.

6.3 CD versus Telephone

There are different levels of fidelity in recordings. CDs are popular because they provide an accurate reproduction of the original sound, or at least those parts that can be heard by humans.

Because humans (young ones anyway) can hear up to 20 KHz, the sampling must be at least 40 KHz. CDs are sampled at 44,100 Hz, providing clear fidelity beyond the limits of human perception.

Telephone speech is not as clear as a CD. In fact, it is very difficult to distinguish between the letters "s" and "f" over a telephone. (Try it.) This is because telephone speech is typically sampled at 8000 Hz, and filtered to a resulting maximum frequency of about 3300 Hz. All parts of the speech signal that are beyond 3300 Hz are absent from the telephone signal that is heard by the listener.

Here are several common digital media and rates:

Medium	Rate Hz
Digital Audio Tape (DAT)	48,000 Hz
Compact Disc (CD)	44,100 Hz
Radio	16,000 Hz
Telephone	8,000 Hz

6.4 Quantization Error

Sampling rate creates one type of recording error. Another type comes from quantization. When the signal is sampled at 0.002371564 volts, how accurately must we record it? How much does it matter?

For telephone speech, we typically use 8-bit quantization. This means that we divide up the signal strength into $2^8 = 256$ buckets and we assign each sample to the bucket that contains it. There might be a bucket from 0.00236 to 0.00240, and all signals in that range would get the same measurement.

This is similar to asking how tall someone is. Perhaps they will say 5 feet 7 inches. But how accurate is that? Their actual height could be anywhere from 5 feet 6.5 inches to 5 feet 7.5 inches and their answer would still be accurate.

The quantization error can be determined by asking how many bits are in each sample. Eight bits is common for telephone speech. 16 bits is common for CD recordings. A good quality microphone might be able to deliver 13 bits of accuracy.

6.5 File Formats

There are several file formats in which speech data can be saved. In recent years the most popular format (most frequently used format) has been Microsoft WAV format. There are many other formats. The difference is in how many samples occur per second, and exactly each sample is represented, and whether it is stereo (two channel) or mono (one channel), and whether the quanta are logrithmically spaced or spaced in some other way.

Researchers at Cambridge University have developed a file conversion utility named `sox`. The interested stu-

dent is referred to their documentation for further information.

7 Spectrograms

During World War II spectrogram technology was classified by the US government as a critical war technology. A spectrogram was called a voice print, and analysis of captured auditory communications was done using spectrograms to help identify the speaker.

Today it is no longer classified, and it is used for linguistic analysis.

7.1 Fourier Transform

A crucial step in the creation of spectrograms is the use of the Fourier transform. This mathematical tool converts a sound signal into the frequency domain, allowing us to see the component sine-wave signals that make up the original composite signal.

7.2 Windows

The Fourier transform operates on a series of sample points. Typically one might choose 1/100 of a second of telephone speech, comprising 80 samples in a vector. The transform is accomplished by means of a matrix multiplication, resulting in another vector where each element represents a signal component.

To get the best (most stable, reliable, robust) results, the windows should be aligned to pitch marks, but this is often inconvenient.

7.3 Formants

Each window of speech is transformed into frequency strengths, and these can be plotted in a time-versus-frequency plot. Generally the stronger signals are represented as a darker grey, and the weaker signals as a lighter grey. Unfortunately, there is rarely if ever a perfect 100% signal or a perfect 0% absence of signal.

When viewed by a human, the darker regions form bands that move across time. These bands are called formants. They are numbered starting at the lowest frequency, as F_0, F_1, F_2, etc. Each band represents a resonance in the speech production system of the person talking. As plosive consonants (b, p, t, d, g, k) explode, the spectrogram shows the exact moment of the transition from stop (closure) to asperation. As diphthonged vowels (bite, bait, butte, boat, bout, yow) transition across time, their transitions can be viewed in the movement of these formant bands.

8 Phonemes

The International Phonetic Alphabet (IPA) is the standard, accepted orthography for writing phonemes. In fact, English was once a phonetic alphabet, but over time the letter-to-sound correspondence has been stretched almost beyond recognition. The word "phonetic" comes from the ancient Phonecians of the middle East, who developed a simple alphabet of sounds in contrast to the prevailing writing systems that were pictographic. The concept of a phonetic writing system is that the pronunciation of words can be read directly from the graphical representation of those same words. Each symbol has a standard meaning that does not vary.

For computer work, we have adopted WorldBet developed by James Hieronymous of AT&T Bell Labs, which is a convenient alphabet for computer work, and is based on the IPA.

8.1 Production Tolerance

Tolerance is the amount of variation that is allowed before something becomes unusable. When producing things, it is generally good to maintain tight tolerances, so that all things of the same type form a tight cluster in terms of variation.

In the production of speech, the IPA and WorldBet recognize subtile variations and allow for orthographic transcription of these details.

8.2 Perception Tolerance

In the utilization of something produced, it is generally good to provide wide tolerances so that things of the same type from different producers still fall within the same loose cluster in terms of variation.

For purposes of speech recognition, this means that the same word pronounced by different speakers should still be recognized as the same word. The number seven might be pronounced as seven, sevin, sebin, sebun, sevun, sivin, siun, sen, or sebu, and still be recognized by a typical listener.

8.3 Foreign Accent

English carries a distinction between the vowel sounds in "beat" and "bit." Some languages do not. Native speakers of these languages will not discriminate between beat and bit. I believe this is because their natural perceptual tolerance has created a loose category in their mind, and both vowels fall into the same category.

To me this seems similar to musical pitches. On the piano, there are 12 tonal steps from one octave to the next. They are logrithmically spaced, with each frequency almost 6% higher than the one below. But on a slide trombone, there are an infinite number of pitches possible from one octave to the next. What do we hear

when someone plays a pitch between two of the standard piano notes? What do we do when asked to find the note that was played? Most of us have not developed the precision in hearing to recognize such a small difference, but some people can.

As foreign speakers mimic English speakers, they will hear the phonemes to which they are accustomed, and will reproduce them using their own tight production tolerances. This can result in a Spanish speaker saying the word "scissors" the same as an English speaker would say the word "Caesar's."

These vowel modifications are predictable and consistent. Because of this recognizable consistency, native speakers will recognize the foreign speaker as having an accent, or a consistent near-miss pronunciation of many words.

It should be possible to use this consistency in speech recognition systems to identify the accent of the speaker and then use that knowledge to better identify words the speaker is saying.

8.4 Neural Networks

Artificial Neural Networks (ANNs) are computer systems made from collections of artificial neurons. They accept a vector of inputs and produce a vector of outputs. They compute their results in constant time. They mimic what we know about the human nervous system. They are trained by presenting them with sample inputs and corresponding correct outputs, and working to minimize the recognition errors.

In a speech recognition system, each input typically represents one feature of the captured speech signal. The combination of these feature strengths results in an output vector that shows, for example, the likelihood that these inputs represent various phonemes under consideration. For example, a given frame of speech may look 30% like an ee and 12% like an ih, and 0.1% like an s.

8.5 Viterbi Search

Given a phoneme string and a set of neural network outputs, it is possible to measure how well they match. One such algorithm is called Viterbi search. The search results in an optimal alignment between the utterance and the target phonemes. This is true even if the phonemes do not match the utterance. There will always be a score. Fortunately, the score for the correct phonemes will be much higher than the score for incorrect phonemes. Unfortunately the score is never 100%, because it is the result of multiplying the probabilities of each individual phoneme assignment, frame by frame, across the entire utterance. (Actually to speed things up, the result comes from adding the logrithms of the probabilities, which has the same result.)

There are serious drawbacks to using Viterbi search. There is a clear problem with using probabilities from a neural network because the assumption is that the probabilities are disjoint and sum to 1.0, when in fact there is always some correlation between similar phonemes, such as s, f, sh, and v. Second, even if the inputs are perfectly disjoint Viterbi does not take into account the duration of each phoneme. Third, the log probabilities are summed across the frames, but this is also incorrect. Dispite these drawbacks, Viterbi is a simplification of reality that performs remarkably well in the real world. But there is clearly room for improvement.

9 Fluent Speech

Another serious issue in speech recognition is the flowing nature of speech. One word blends into the next, especially when similar sounds come together. Say "six sheep" for instance. Does it sound the same as "sick sheep"? How about "five four." Does it sound the same as "fie four"? Whole phonemes can easily vanish between words.

This creates a serious problem in speech recognition because we cannot easily tell where to divide the phonemes into words. Somehow the human mind does it, but it is hard to program a computer to do it.

Here are some examples from number recognition.

nine	is in	ninety
ninety	is in	nineteen
nineteen	is in	ninety-nine
oh	is in	f-ou-r
one	is in	t-wun-y
two	is in	tw-elve
two	is in	tw-enty
and	is in	h-und-red
sixth	is in	six three

9.1 Word Boundaries

Because of fluency and coarticulation effects, it is frequently impossible to assign a phoneme to one word or the other. It may need to be shared between two words. In rare cases, a string of several phonemes may be shared between adjacent words in fluent speech.

9.2 Mondegreens

Ambiguity can result from conflicting explanations of the phonemes that were perceived. Some of the most interesting examples come from song lyrics that are incorrectly heard. Here are some examples:

```
they hae slain the earl o morey
  and lady mondegreen
they hae slain the earl o morey
  and laid him on the green
```

```
'scuse me while I kiss this guy
'scuse me while I kiss the sky

olive, the other reindeer
all of the other reindeer
```

Based on Lady Mondegreen and her appearance in a magazine article many years ago, these constructs are often called Mondegreens. An Internet search will turn up many interesting examples.

9.3 Resolving Ambiguity

Somehow the human mind resolves ambiguity in many cases without even noticing it. The most likey result is chosen and the other contenders are discarded without a thought. The likelihood probably depends on the context in which the utterance is heard.

Here is a common phrase that has many possible interpretations, but only one that is generally recognized.

```
time flies like an arrow
```

Usually this is understood to mean that time goes by quickly, and we should seize the day before it escapes us.

It could also be a discussion of the eating habits of flies. Fruit flies like a banana. What do time flies like?

It could also be an instruction for contest judging. We have been timing arrows. Now our task is to judge a race among flies. But how do we time flies?

A computer program using a dictionary of meanings, parts of speech, and English grammar might come up with all three of these interpretations and perhaps more. Additional information is needed to identify the correct meaning.

9.4 Common Sense

Based on my laziness conjecture, humans articulate their speech carefully enough to be understood by other humans. "juwanna eat" is recognized at lunch time, but might be misunderstood when pointing at an airplane. Because of this laziness in speaking, and the need for additional information to resolve ambiguity, automatic speech recognition seems likely to remain a difficult task.

There are some interesting efforts under way to help computers gain the common sense that humans take for granted. One of the most notable efforts in this area is Douglas Lenat's Cyc project, which offers free common sense data on the Internet.

10 Understanding

Understanding is the reduction of all these speech inputs into a completed perception.

When given a list of words, and asked to identify any that were repeated, subjects did not confuse stone and bone, but they did confuse stone and rock. Why is that? Apparently our perception is not tied to the phonemes that represent the object. There is something else going on.

11 Response

In human interaction, we are expected to react to incoming speech. Perhaps we move in response to a warning. Perhaps we mutter agreement. Perhaps we ask for clarification. Perhaps our incorrect behavior causes the speaker to assume a misunderstanding and to clarify their original intent. Carrying on a conversation with a human is much more robust than the fragile activity of computer speech recognition.

Reminder

As I mentioned above, this tutorial is presented as an overview and as a study guide for an undergraduate elective course in Automatic Speech Recognition, CS 441, taught at Brigham Young University Hawaii by Don Colton. The document is in an early stage of development and includes many rough-hewn statements that deserve and anticipate more accuracy and supporting citations in future drafts. In particular, it contains statements that I have heard and/or believe to be true, but which I have **not yet verified.** I hope to verify and expand on many of these topics over time, but this is what I have available today.