Speech Processing Background

Sam Roweis

November 1998

1 Introduction

This note provides an extremely brief and necessarily incomplete introduction to speech processing by machines for those unfamiliar with the basics of the field. It is clearly beyond the scope of such a tutorial to give a comprehensive survey of computer speech processing methods. Below I provide a very general overview of the current paradigms used in speech processing. I do not, however, provide details of implementing a recognition system; although unfortunately most of the work in getting a system to actually function properly is in the details. I also give examples of state of the art performance for recognition, synthesis, speaker identification and compression systems. Three standard textbooks, one old but classic (by Rabiner and Schafer [42]) and two newer (by Rabiner and Juang [41] and Deller , Proakis and Hansen [14]), provide very comprehensive introductions to this material. The collection edited by Waibel [48] provides an excellent source of important early papers. Several very good books have been written on speech science in general with less of an engineering focus (see for example [24, 21, 22, 29, 25, 26, 40, 37, 18]).

2 A brief historical review

The history of modern speech processing genuinely begins after the second World War. Systems recognizable as the direct predecessors of current systems appeared in the early 1970's. A brief historical review[1] for cultural as much as academic benefit follows:

- **1875** Edison invents the **phonograph**.
- **1876** Bell's **invention of the telephone** [7, 8] inspired by the attempts of Sir Charles Wheatstone to reproduce the speaking machine of Wilhelm Von Kempelen [20].
- **1913** Markov models [35].

[1] Flanagan's book [24] is the source of many of these items. Thanks also to Nelson Morgan for some items and for the reference to Radio Rex.

- **1915** The first **completely automated transcontinental telephone call** was made from New York to San Francisco ushering in the age of **universal service** in which any telephone user could connect themselves instantly to any other user. By the beginning of the first World War it had become clear that telecommunications would have an enormous impact on society.

- **1922 Radio Rex** toy dog. The first device I am aware of to use speech recognition. A description from David and Selfridge: "It consisted of a celluloid dog with an iron base held within its house by an electromagnet against the force of a spring. Current energizing the magnet flowed through a metal bar which was arranged to form a bridge with two supporting members. This bridge was sensitive to 500Hzacoustic energy, interrupting the current and releasing the dog. The energy around 500Hzcontained in the vowel of the word *Rex* was sufficient to trigger the device when the dog's name was called." [12].

- **1934** The Sherman Antitrust Act and later the **Communications Act** were passed in response to the predatory growth of monopolies such as Western Union and later AT&T. The Communications Act of 1934 created the Federal Communications Commission (FCC) whose mandate was "regulating interstate and foreign commerce in communication by wire and radio so as to make available, so far as possible, to all the people of the U.S. a rapid, efficient, nationwide and worldwide wire and radio communication service..." [11], in other words it made telephone service a *right*. This brought telecommunications technology into the American home for good.

- **1938** The voder and vocoder were invented by Dudley [20] and found important use providing secure voice communications between Roosevelt and Churchill during World War II.

- **1946** ENIAC, an early digital computer.

- **1946** Invention of the **sound spectrograph** [32].

- **1951** A simple **phoneme detector** [44].

- **1952** The **Audry** spoken digit recognizer [13] using formants. [2]

- **1954** Baumann word recognizer [6].

- **1956** Olson and Belar (RCA) spoken digit recognizer [36].

- **1957** Bellman's book on **dynamic programming** [9].

- **1958** Wiren-Strubs attempt to use **linguistic features** such as voicing, turbulence, and onset times.

- **1958** Dudley uses a continuous spectrum evaluation rather than formant or other feature tracking [19].

[2] In the early 1950's, driven by telephone company interests, there were several research groups trying to build devices that could recognize spoken digits. Computers were just being invented, and so many systems were still analogue electronics. Davis et al. demonstrated *Audry*, a single speaker isolated digit system in 1952. Audry worked by finding formants and had a surprisingly low error rate of 2%.

- **1959** Denes, Mathews and Fry phoneme recognizer which used derivatives of spectral energies as well. Denes also introduces a simple bigram **language model** for phonemes which is the first example of non-acoustic information in recognition [28, 16, 17].

- **1962** Review paper in IRE by David and Selfridge [12].

- **1968 Dynamic time warp** first applied to speech by Russian engineer Vintsyuk [46, 47]. Also later by Itakura and Sakoe & Chiba [43].

- **1969 Pierce's caustic letter** criticising speech processing engineers for unprincipled hackery [39].

- **1966-72** Many important theoretical advances in the study of **probabilistic functions of Markov chains** (which later came to be called **hidden Markov models**) by Baum and colleagues at IDA [4, 3, 5, 2]. Largely unknown in the speech community.

- **1968 Linear predictive coding** (LPC) methods first applied to speech parameterization by Itakura, Atal & Shroeder, Markel and others [1, 34, 33].

- **1971-76** First **ARPA** project for US$15M involving CMU, BBN, SDC, Lincoln Labs, SRI and Berkeley. Resulted in the **Harpy** and **Dragon** recognition systems.

- **1975** John Ferguson at IDA, Jim and Janet Baker at CMU (now Dragon) and Fred Jelinek's group at IBM (now Hopkins) all begin **applying HMMs to speech recognition**.

- **1977** Dempster, Laird and Rubin recognize the general form of the Baum-Welch updates and name it the **EM algorithm** [15].

- **1980** A landmark symposium in Princeton hosted by IDA reviewed current work up till that point [23].

- **1985-88 Standard corpora** begin to be used across research groups. The **TIMIT** database is collected by Texas Instruments and NIST.

- **1986-7** The second ARPA project involving the **Resource Management** (RM),**Wall Street Journal** (WSJ) and **air travel** (ATIS) databases.

- **1988-90** New features such as mel-cepstra (Bridle), PLP (Hermansky), and delta/delta-delta coefficients (Furui) improve the performance of systems on larger and noisier tasks.

- **1990s** Neural nets successfully used as output models in large HMM systems [10].

- **1990s** The popular Cambridge **Hidden Markov Model Toolkit** (HTK) helped to standardize research code and increase access to new players in the field.

- **1990s** Discriminative training, vocal tract normalization, speaker adaptation.

3 Basics

3.1 What is speech?

Speech is the generic name given to sounds which carry language content. It often has a physical manifestation as a longitudinal compression wave, but for our purposes a speech signal will be a one-dimensional time varying signal. Furthermore, we will make the assumptions that the signal is band-limited, i.e. all of the useful information has its power in a finite frequency range, and that our receivers (ears or synthetic microphones) have a finite acuity. This means that we can *sample* the original time signal which is real and continuous and use instead a vector of values discretized both in time and amplitude. This definition of speech freely admits that we are ignoring many effects such as binaural hearing and other modalities, for example lip reading that are known to be important in human perception; on the other hand humans can speak and understand speech perfectly well over the telephone lines—a situation in which all these assumptions are directly and literally enforced.

3.2 Is the time domain representation best?

Above reference was made to "the speech signal", but it was not specified which combination of physical variables are measured to get that signal. The most direct parameter to measure is the air pressure of the longitudinal wave created when a person speaks. This is informally known as the intensity waveform and all microphones measure something roughly proportional to this or to its first derivative by transducing the movement of the air into an electrical voltage using their diaphragms. However, each individual microphone exhibits a slightly different and usually mildly nonlinear transformation from the true pressure signal to its output voltage. Furthermore, ambient noise from other sources and effects of the medium in which the sound is traveling mean that the pressure wave at the microphone is never the same as that which left the speaker's mouth. All this means that from the outset we can never hope to have access to the exact pressure waveform produced by a talker. Nonetheless, microphone intensity signals definitely contain an enormous amount of structure—after all we can record and play back speech that sounds just fine. However, this structure is extremely difficult to extract from the mass of information in the intensity waveform. Consider the waveforms in figure 1, both of which are myself saying my name and which were recorded on different days but using the same microphone in the same room. These simple figures show that there is quite a lot of *variability* in the raw speech waveform even across the same person saying their name (something they say quite a lot) twice using the same microphone and in the same acoustical environment.

This difficulty motivates us to search for some transformation of the raw intensity waveform into a different representation where the important structure is easier to identify and the enormous amount of variability is reduced. The hope is that a new representation will make it easier to do whatever tasks we are interested in be it speech recognition, speaker recognition or even synthesis and compression.

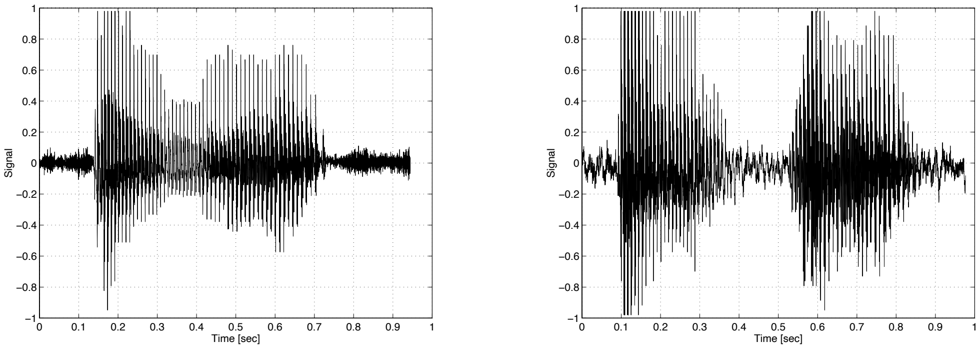

Figure 1: The author saying his name twice. Notice that the time domain waveforms appear quite different even though the same putative sounds are being spoken by the same speaker using the same microphone in the same acoustical environment.

3.3 A simplified model of speech production: excitation plus filtering

Our hope is to extract features of the speech signal that make our tasks easier. A few simple observations about how speech is produced will motivate the features commonly used in speech processing. The first is the relatively basic observation that speech is sometimes harmonic or "song-like" and sometimes noise-like or "breathy." These two modes relate to whether or not an oscillation[3] of the vocal cords is modulating airflow from the lungs and whether or not small constrictions in the airflow path are causing turbulent noise sources. They are called the *voiced* and *unvoiced* excitation modes of speech. Furthermore, in the voiced model there is the simple idea that some people have low pitched voices and others have high pitched voices. Loosely speaking, pitch is related to the frequency of the oscillations which drive the rest of the vocal system. In *atonal* languages like English pitch does not carry information about phoneme identity, although it does carry prosodic information about questions or emphasis. Another important observation is that when a person speaks, their vocal tract and articulators form into a particular sequence of configurations that give the sounds of different words their different spectral character. Again loosely, this can be thought of as related to the "modulation" or "shaping" of the excitation. A final observation is that the signal power in speech varies significantly over time and that the timing of these various modulations of pitch, energy and spectral shape is different from word to word or syllable to syllable and is perceptually very important.

These observations lead us to a bare-bones model in which we treat speech sounds as though they were produced by either a periodic pulse generator or a white noise source (the vibrations of the vocal cords for voiced speech or the turbulent air flow for unvoiced speech) that was filtered by a simple linear filter (the tube shaped by the articulators). Since the spectrum of the sources are more or less broadband, the spectrum of the produced speech should be the transfer function of the filter. Since the filter is

[3] The oscillation of the vocal cords is a complicated nonlinear phenomenon called *Bernoulli oscillation.* The elastic force in the cords works to close the vocal folds, while the airflow from the lungs below can force them open. As the airflow speed increases, the pressure drop between the folds causes them to snap shut. Pressure then builds up until they are forced open again.

determined by the mouth and articulators, if we can capture the transfer function of the filter in this model, then we will have captured the state of the articulators.

3.4 Useful features for speech: spectrograms

The simplified model above suggests that if we could find a way to estimate the voicing (source characteristics), the shape of the vocal tract and articulators (modulator characteristics) and signal energy during speech, then we would be in relatively good shape. One more key point to make here is that since the vocal tract and voicing parameters are changing quickly over time as the person speaks, what is needed is a *time-varying* parametric description of these properties and not a single set of parameters for the entire signal. Such a time varying representation is typically achieved by breaking up the signal into segments, called *analysis windows* or *analysis frames* which are short enough that the parameters of interest can be assumed constant within each one. Parameters are then estimated within a window which slides across the signal to obtain a time-varying sequence of parameters.

There are many possible signal features to consider – how should we choose between them? Ideally, we would like to select based on which features most reliably capture the articulator and voicing parameters of the signal over time. However, since all we have is the acoustical signal, we do not know what those parameters really are, and thus it is difficult to judge how well a given transform is preserving them. Instead, we can study certain mathematical properties of the acoustical signal which we belive capture in a very crude way some information about the shape of the vocal tract and the behaviour of the source over time. Then we can compare features to see how well they estimate these mathematical properties.

One obvious set of features to consider given the above are the *short-time magnitude spectra* of the signal. The magnitude spectrum tells us which frequencies contain the energy in our signal. The shape of the spectrum of a short window of speech is related to the articulator parameters through the simple models of speech production described above. Am image of such a sequence of short-time spectral features is generically known as a *spectrogram*. The name comes from a traditional method of displaying this information, illustrated in figure 2. A spectrogram shows signal information in the time-frequency plane. Time runs across the horizontal axis, frequency up the vertical axis. The amount of energy in any region in the time-frequency plane is indicated by the darkness of shading. Such a display is constructed by taking short windows of the original time domain signal and performing (windowed) spectral analysis on each one; the resulting vectors are aligned vertically and stacked one to the right of the previous. This involves taking the discrete-time Fourier transform of each analysis window. It is usual to apply a windowing (such as a Hamming window) to reduce edge effects and also to restrict the length of the frames so that the number of samples is a power of two. These retrictions mean we can use the fast Fourier transform (FFT) rather than a slower generic DFT algorithm.

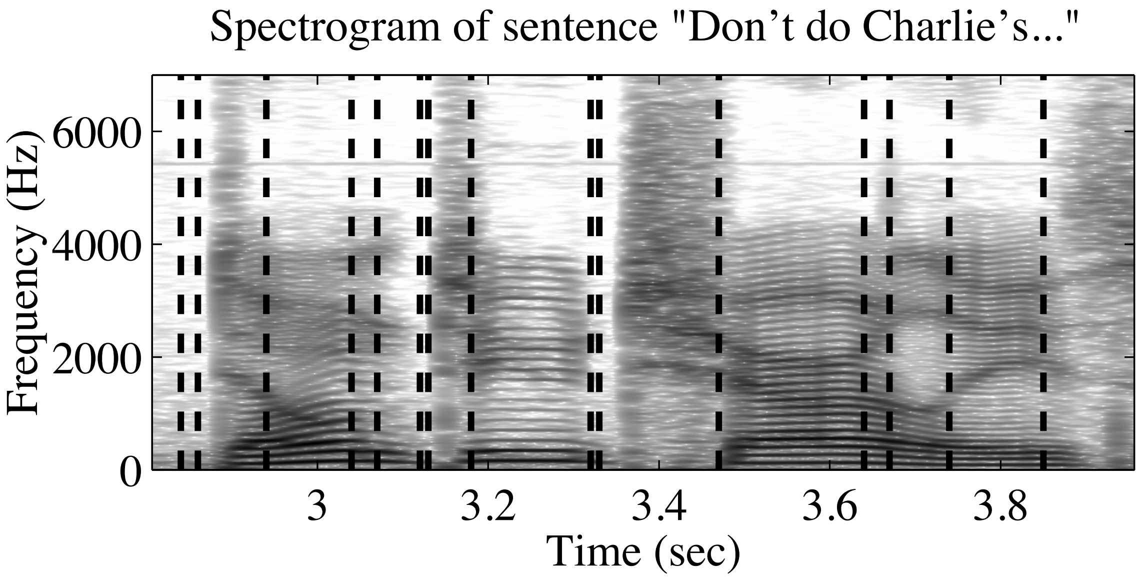

Figure 2: A spectrogram showing signal information in the time-frequency plane. Time runs across the horizontal axis, frequency up the vertical axis. The amount of energy at any point in the time-frequency plane is indicated by the darkness of shading. Such a display is constructed by taking short windows of the original time domain signal and performing (windowed) spectral analysis on each one; the resulting vectors are aligned vertically and stacked one to the right of the previous. In this example, the vertical bars show the rough segmentation into phonetic units as determined by a forced-alignment procedure.

3.5 Smoothed-spectrogram features: linear prediction and cepstral coefficients

Unfortunately it is often not enough to just take a speech signal, break it up into analysis windows, and take the discrete-time Fourier transform of each one. One reason is that the short-time Fourier transform of a small speech window gives a very noisy estimate of the power at any frequency. We need to find a representation that captures the essential shape of the spectrum and rejects the noise so that we can compare two spectra sensibly. The other reason is that the full short-time spectrum, even if it were not noisy, still contains just as much information as the original intensity signal; and that is too much information for us to process. We need to find a more compact representation of the spectrum. These two ideas are really one and the same: the spectra are typically *smooth* and so we want to filter our noisy estimates and represent them using fewer coefficients.

How can we go about reducing the amount of information used to represent each speech frame while at the same time preserving the basic spectral shape? We require a definition of spectral distance to compare the original spectral shape with the shape implied by the reduced representation. This can be quantified in several ways. One is to consider the *residual energy* between the two spectra. This is the integrated square difference between the *linear* spectra. Because of Parserval's relation this residual is proportional to the energy difference between the original signal and our reconstruction of the (time domain) signal based on whatever reduced representation we select. This difference measure between spectra leads us to a set of coefficients that are optimal in the sense that they minimize the squared difference under certain linear constraints. For minimizing squared error on the linear spectrum or in the time domain, we are led to the *linear predictive coding* (LPC) coefficients and closely the related *cepstral coefficients*. Note that inherent in our distance measure is the assumption that only the *magnitude* of the spectrum is important, and not the *phase*. For short time windows (less than about 40ms) this is certainly true perceptually: a signal reconstructed from only short time magnitude spectra sounds much like its original (it is intelligible) than one reconstructed from only phase spectra (which just sounds like noise); furthermore, humans are insensitive to many manipulations or randomizations of short time phase.

3.6 Fixed vs. signal dependent transforms

Before we examine the two important transforms mentioned above, it is important to make a brief point about the idea of *fixed* versus *signal dependent* transforms. A fixed transform is one in which the basis vectors of the new (transformed) space do not depend on the original signal. In fixed transforms, if a sender and a receiver agree on the type of transform beforehand, then the sender can take the original signal, transform it, and send some or all of the coefficients to the receiver. The receiver can then reconstruct the signal based solely on the values of the coefficients. However, with *signal dependent* transforms, the sender must first communicate the basis vectors to be used for reconstruction in addition to the coefficients. Here we are considering fixed transforms. But it is worth remarking that if we were to permit the transforms to be signal dependent, then the *Karhunen-Loeve Transform* (also known as Principal Component Analysis) would be optimal in terms of residual energy. This means that the best we could ever do for

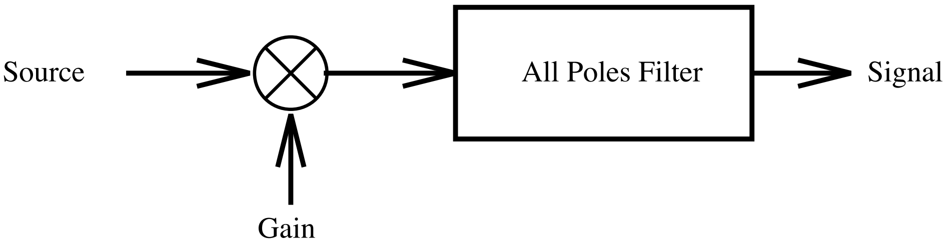

Figure 3: Linear predictive coding model. Each sample of an incoming signal is modeled as a linear combination of some fixed number of past samples plus a driving noise or "innovation" term.

minimum squared reconstruction error would be to project our speech signal onto the first few principal components of the distribution from which the signals were drawn. However, we are investigating here only fixed transform approaches. (Although it is true that a fixed PCA basis could be chosen by prior training on a large databse.)

3.7 Linear predictive coding (LPC)

Linear predictive coding (LPC) [34, 33] is a model of a discrete-time signal s_t in which we try to approximate each sample by a linear combination of some previous samples:

$$s_t \approx a_1 s_{t-1} + a_2 s_{t-2} + a_3 s_{t-3} + \ldots \tag{1}$$

The above approximation is known as an *autoregressive* (AR) signal model. The number of previous samples on which the current sample depends is the *order* of the model. We make up the difference between this approximation and the actual signal by a scalable correction function u_t called the *source* or *innovation* function:

$$s_t = a_1 s_{t-1} + a_2 s_{t-2} + \ldots + a_k s_{t-k} + G \cdot u_t \tag{2}$$

If the source function could magically take on the negative of our prediction error at any time, then the equality above would indeed be exact. We can think of the source function as the input to a system and the actual signal as the output. What kind of system lies in between? Let us take the *z-transform*:

$$S(z) = a_1 \frac{S(z)}{z} + a_2 \frac{S(z)}{z^2} + \ldots + a_k \frac{S(z)}{z^k} + G \cdot U(z) \tag{3}$$

Thus the transfer function from source u_t to output s_t is known as an *all-poles filter* whose impulse response are the coefficients a_k and whose frequency response is given by:

$$H(z) = \frac{S(z)}{G \cdot U(z)} = \frac{1}{1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_k z^{-k}} = \frac{1}{A(z)} \tag{4}$$

Figure 3 shows a block diagram of this model.

For speech signals, the assumption of an all-pole filter is not a bad one because of the nature of the process that generates speech.[4] Thus, the source function need only be a

[4] Strictly speaking, if we define the "transfer function" of the vocal tract to be the ratio of volume velocity at the output to volume velocity at the input then there physically cannot be any zeros. However, it is important to acknowledge that it still may be empirically useful to model the observed spectra using a small number of both poles and zeros to give a better fit.

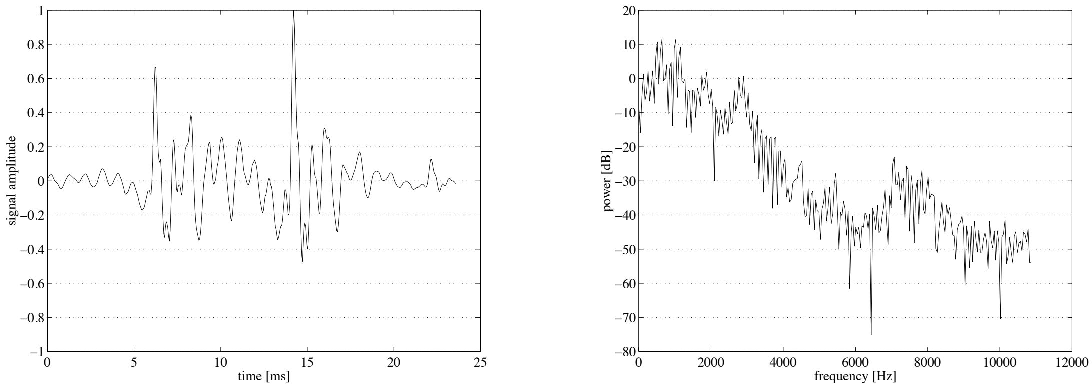

Figure 4: A short window of speech (left) and its corresponding spectrum (right). The speech has been windowed with a Hamming window to reduce edge effects in the spectral analysis.

relatively small amplitude “noise” source, representing the small and random prediction errors. Following our simple model of speech production, we define the source as either an impulse train for voiced speech segments or white noise for unvoiced speech. Typically only a few poles are required in the filter since the usefulness of far away (long ago) samples for predicting the current sample point is quite small. A good rule of thumb is that samples up to 1.5ms in the past help in predicting the current sample, but not earlier.

To see how this works, at least to some degree, consider a concrete example: In figure 4, I have shown a window of a speech signal that is 512 samples long (which at the sampling rate of 22kHz is about 25ms). It has been multiplied by a Hamming window. Its spectrum is also shown.

Figure 5 shows the same speech sample with the spectrum reconstructed by an LPC model of order 12. To do this, I have estimated the best 12 pole filter, kept only the 12 coefficients $a_1 \dots a_{12}$, and reconstructed the spectrum based on these 12 numbers.

Linear predictive coding coefficients are used extensively in academic and commercial speech systems. The other extremely common features are the cepstral coefficients (introduced in the next section) which in some cases give superior performance and can be mathematically easier to work with.

3.8 Cepstral coefficients

The cepstrum[5] transform (see for example [30]), while having some complicated associated mathematics is really based on a very simple idea: the short time log magnitude spectra that we are trying to capture are quite noisy and we only want to capture their essential shape. So let us get rid of the noise by “low-pass” filtering that shape. Low-pass in which sense? In the sense that we only want to let the log magnitude spectrum change

[5] The name *cepstrum* as well as other terms from the literature such as *liftering* and *quefrency* come from the amusing reversal of the beginning of the words *spectrum*, *filtering*, and such. This is because the cepstrum, as we will see, is an inverse transform of a transform.

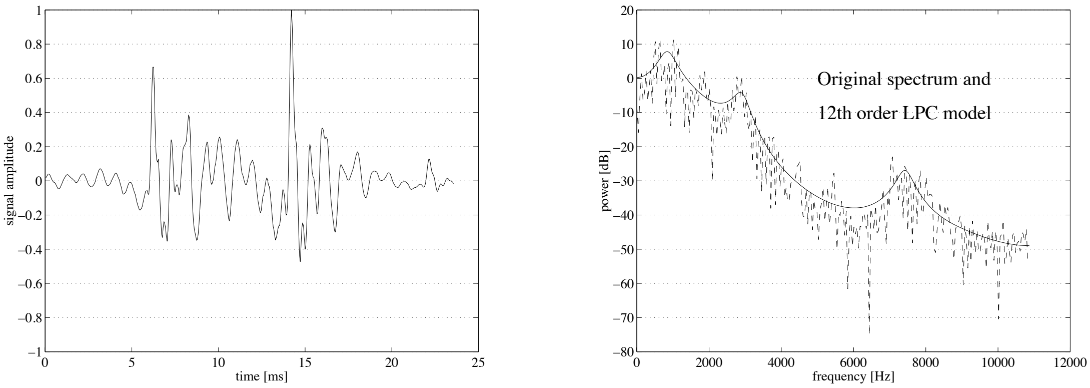

Figure 5: The segment of speech from figure 4. The right panel shows the spectrum obtained by linear predictive coding after retaining only a few coefficients. A 12^{th} order LPC model was used. The spectral fit is quite good in terms of residual energy. Compare with figure 6.

slowly across frequency. That is, we want to eliminate large changes in log magnitude over small frequency intervals. What we want to do is "smooth out" the spectrum in exactly the domain as one sees it in all the figures above; namely log magnitude.

In order to achieve this, we do what at first appears to be a crazy thing: *we take the spectrum of a spectrum.* That is, we take our original signal, compute its log magnitude spectrum and then take the spectrum of *that*, chop off all the upper coefficients (i.e. "low-pass" filter) and then inverse transform to reconstruct our original log magnitude spectrum. The domain in which we are when we chop coefficients is called the *cepstrum.* More formally, the cepstrum of a real signal $s(t)$ is defined as:

$$C(q) = F^{-1}\left(log|F\left(s(t)\right)|\right) \tag{5}$$

where $F(\cdot)$ denotes the discrete-time Fourier transform and $F^{-1}(\cdot)$ denotes its inverse. The variable q denotes the axis of the cepstral domain, called *quefrency*; which is something like the rate of change over frequency analogous to regular frequency ω being the rate of change over time. Notice that the cepstrum is real and also symmetric in q since the log of the magnitude spectrum is real. Finally, notice that we have taken the inverse transform F^{-1} after the log, and not the forward transform, but they differ only by a scaling factor. The cepstrum is essentially trying to approximate the log magnitude spectrum using a truncated Fourier series:

$$log|F(\omega)| \approx \sum_{k=-K}^{K} c_k e^{-jk\omega} = log|F'(\omega)|$$

where $c_k = c_{-k}$ are the real cepstral coefficients and $F'(\omega)$ is the cepstral approximation to the spectrum. Thus we are smoothing the log magnitude spectrum.

The hope of the cepstrum transform is that in the cepstral domain there will be some power at low coefficients representing the slowly varying (in frequency) component of the spectrum and some power at high coefficients representing spectral noise (and the fine spectral structure due to voicing). If we only keep the lowest few components and set

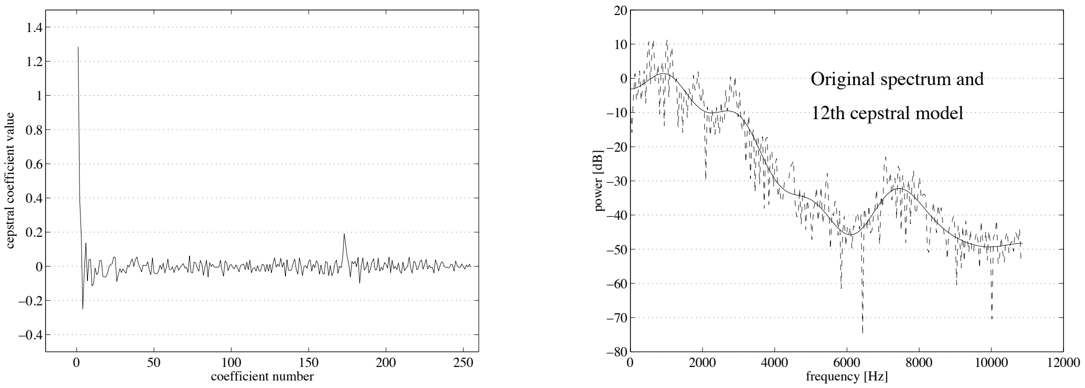

Figure 6: The coefficients resulting from cepstral analysis on the segment of speech from figure 4. Notice the smaller second spike at higher coefficient values. The right panel shows the spectrum obtained by keeping 12 cepstral coefficients and reconstructing.

all the rest to zero, then when we reconstruct the log magnitude spectrum it should be "smoother." This truncation process (or more generally a windowing process in which higher cepstral coefficients are de-emphasized) is called *liftering.* Shown in figure 6 are the cepstral coefficients for the speech segment of figure 4 and the reconstructed spectrum obtained by keeping only the first 12 cepstral coefficients and setting the rest to zero.

One can see that the cepstral coefficients have a bump at low quefrency and another bump at high quefrency. The high quefrency bump is the pitch excitation of the voiced speech. One can also see that the reconstructed spectrum satisfies the original goal of being "smoothed" in the log magnitude domain. Notice that this reconstruction is technically "worse" than the LPC reconstruction in terms of residual energy but captures the essential character of the spectrum just as well (compare with figure 5). The cepstral coefficients are also easier and faster to compute in some cases.

3.9 Spectrum-scale perceptual effects: mel and Bark scales

Why did we use a logarithmic scale for frequency in working with the cepstrum above? The reason is partly one of engineering history but also partly that this scale has been found to be perceptually more relevant than the linear frequency scale. Two important effects from human perceptual studies motivate this claim. The first observation is that humans are not sensitive to frequency information below about 200Hz or above about 16kHz for many speech processing tasks including speaker recognition. The upper limit does not typically help us since our Nyquist rate is normally far below it. But the lower limit suggests that we might be able to high-pass all of our short time spectra above 200Hz before computing our cepstral coefficients. This would save computation and possibly add some robustness against external noise sources (fans, 60 cycle hum, etc.). The second observation is that humans pay relatively more attention to the lower frequencies than the higher ones. Roughly, frequencies up to 1kHz are "weighted" in importance linearly while above about 1kHz the perceptual importance falls off logarithmically for a wide variety of psychophysical tasks and measurements. This suggests that we should "remap" our spectra to a more perceptual frequency scale before extracting coefficients. Many such

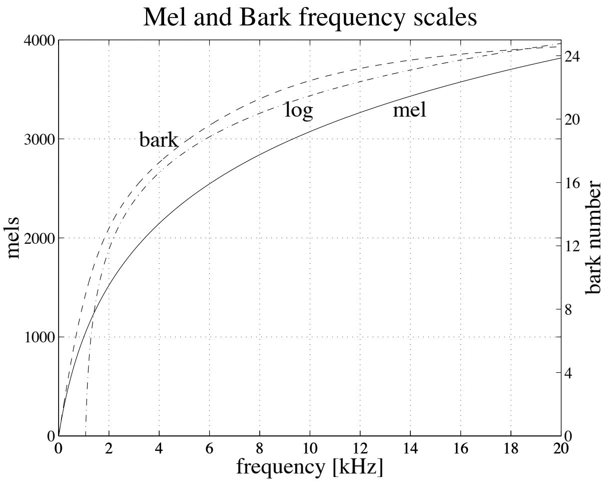

Figure 7: The *mel* and *Bark* perceptual frequency scales. The solid line shows the mel scale, the dashed line shows the Bark scale and the dotted line shows a logarithmic shape for comparison. These scales are all fits to psychophysical curves obtained in pitch matching experiments.

perceptual scales have been proposed, the two most common of which are the *mel* [45] and *Bark* (named after the German acoustician Barkhausen) scales which are based on the results of psychophysical pitch matching and critical band experiments. The scales can be defined by simple mathematical approximations to the measured psychophysical curves. For example, the following transformations (all frequencies are in Hz) are common:

$$f_{mel} = 2595 \log_{10}(1 + f/700) \tag{6}$$

$$f_{Bark} = 13 \arctan(0.76 f/1000) + 3.5 \arctan((f/7500)^2) \tag{7}$$

These relations are just fits to human psychophysical curves; however, researchers consistently report significant improvements using the Bark and mel scales rather than linear scales. When the mel scale warping is used prior to extracting the cepstral coefficients, the resulting features are known as *mel cepstra*. Figure 7 shows the shape of these two perceptual scales compared with logarithmic and linear shapes.

Another very common signal processing step in speech analysis is to first pass the raw sampled speech signals through a *pre-emphasis* filter. This is a high pass system which flattens the spectrum. The main motivation for it is that speech typically has a $1/f^2$ amplitude spectrum which introduces some *spectral tilt* and we want to flatten that out. It has been found to improve the robustness of recognition tasks and does not take very much time. Typically a time domain smoothing rule

$$s(n) \Leftarrow s(n) - \alpha s(n-1) \tag{8}$$

is used, corresponding to a discrete-time transfer function of

$$H(z) = 1 - \alpha/z \tag{9}$$

which is a single pole high pass filter. A usual value for α is 0.95 which boots power at the Nyquist frequency by 32dB over DC power—quite a significant effect.

3.10 Distance measures between spectral features

Central to any pattern recognition application is the notion of the *distance* between two feature vectors. A common choice is the L^2 norm or Euclidean distance, namely the sum of squared differences between corresponding components of the two vectors. However, this is sometimes not a very meaningful distance measure. In the present case we are interested specifically in two kinds of vectors: vectors of LPC coefficients and vectors of cepstral coefficients. What is an appropriate measure of "distance" between two LPC vectors or two cepstral vectors?

The LPC coefficients were attempting to minimize the residual energy between the true (linear) magnitude spectrum of the speech frame and the LPC spectral approximation. Thus it seems reasonable to compare two LPC vectors by comparing the residual energies between each of their reconstructed spectra and some "true" spectrum. Specifically, it can be shown (see for example page 118, Exercise 3.6 in [41]) that the residual energy between the reconstructed spectrum of an LPC vector $\mathbf{a}$ and the true spectrum is given by:

$$E = \mathbf{a}^T \mathbf{R}_x \mathbf{a} \tag{10}$$

where $\mathbf{R}_x$ is a Topelitz matrix formed from the autocorrelation sequence of the true signal x. Thus, the "distance" between two LPC vectors is the ratio of their residual energies:

$$D(\mathbf{a}, \mathbf{b}) = \frac{E(\mathbf{a})}{E(\mathbf{b})} = \frac{\mathbf{b}^T \mathbf{R}_x \mathbf{b}}{\mathbf{a}^T \mathbf{R}_x \mathbf{a}} \tag{11}$$

If we choose $\mathbf{a}$ to be the LPC vector computed on the true signal x, then $E(\mathbf{a})$ is minimized, and hence the distance above will always be greater than unity. This metric was originally proposed by Itakura and Saito and is sometimes known as the Itakura-Saito distortion between two LPC vectors. It is mathematically complex and can be difficult to understand throughly, but I have presented here the intuitive idea behind it.

Following our distance measure between LPC vectors, we can compare two cepstral vectors by comparing the residual energy between each of their reconstructed spectra and some "true" spectrum. As noted above, the cepstrum is essentially trying to approximate the log magnitude spectrum using a truncated Fourier series:

$$log|F(\omega)| \approx \sum_{k=-K}^{K} c_k e^{-jk\omega} = log|F'(\omega)| \tag{12}$$

where $c_k = c_{-k}$ are the real cepstral coefficients and $F'(\omega)$ is the cepstral approximation to the spectrum. Imagine that we expressed the true spectrum by its full (infinite) Fourier series with coefficients a_k. Then by Parseval's theorem, we can write the residual energy as:

$$E = \frac{1}{2\pi} \int_{\omega} |logF(\omega) - logF'(\omega)| d\omega \tag{13}$$

$$= \sum_{k=-\infty}^{\infty} (c_k - a_k)^2 \tag{14}$$

where c_k are defined to be zero outside $-K \dots K$. Similarly, for another cepstral vector b_k, the residual energy between its reconstructed log spectrum and the true log spectrum would be:

$$E = \sum_{k=-\infty}^{\infty} (b_k - a_k)^2 \tag{15}$$

Thus the difference between the residuals of b_k and c_k is given by:

$$D(b, c) = \sum_{k=-\infty}^{\infty} (b_k - c_k)^2 \tag{16}$$

which is just the normal Euclidean distance that we are used to. It is interesting to note that the distance between b_k and c_k does not depend at all on what the "true" spectrum a_k was. These two facts can make the cepstral vectors easier to work with and to visualize than the LPC vectors in which our intuitive sense of squared norm distance does not properly apply.

4 Spectral pattern matching

4.1 Pattern matching for static sounds: spectral peak locations

An influential study in the early history of speech processing was conducted by Peterson and Barney [38] who collected statistics on 76 speakers saying various steady state vowels at the 1939 World's Fair in New York. They noted the frequency and amplitude of the first three formants. When plotted in various ways (for example F2 vs. F3), the different vowels fall into distinct regions of the formant space. This encouraged researchers because it implied that one only had to accurately estimate formant frequencies in order to classify vowels. While this is true for steady state sounds, it was soon discovered (for example see [31]) that during the production of continuous speech the formant frequencies of the various vowels move considerably and depend on context. These *co-articulation* effects led researchers to investigate dynamic pattern matching techniques which attempt to perform pattern recognition on the time evolution of spectral information rather than on a frame by frame basis.

4.2 Pattern matching for sequences of features: dynamic time warping (DTW), hidden Markov models (HMMs) and others

Motivated by the realization that steady state formant measurements were not representative of continuous speech, researchers began looking at the spectrogram representations of words. Amazingly, it was found that after condsiderable practice people could be trained to "read" spectrograms. (See for example the book [40], as well as the many classes on the subject taught by Victor Zue at MIT in the 1980's and 1990's.) This led to a set of pattern classification algorithms which essentially tried to match a spectrogram template to the incoming spectrogram.

Two fundamental difficulties have to be overcome for this simple approach to work at all. The first is the problem of *time-warp.* Although the spectral patterns of many instances of the same word often look similar, they do not arrive at the same rate – in some cases people will say a word faster or slower than in others. Nor is the speeding up or slowing down always uniform across the word. Certain parts of the word may be selectively sped up or slowed down. An elegant solution to this problem was found in the form of the *dynamic time warping* (DTW) algorithm. DTW is an application of Bellman's dynamic programming [9] to spectrogram matching. It provides a method by which an incoming spectrogram may be locally stretched and squished (in time) to optimally match a template spectrogram (see [46, 47, 43]).

The second difficulty is that the spectral energy patterns in multiple instances of the same word exhibit some variability across frequency. Occasionally spectral energy appears in certain frequency bands in which power is normally absent for a given word. Conversely, spectral energy is sometimes lacking in frequency bands that normally contain power. Researchers tried to model this variability using a *probabilistic* spectrogram template. Again, an elegant solution combining the probabilistic template in frequency with dynamic time warping was found in the application of *hidden Markov models* [4, 3, 5, 2, 41, 14] to speech analysis.

4.3 Other acoustic cues: pitch, voicing

Short-time spectral shape and signal energy are the features captured by the spectrogram and smoothed-spectrogram representations of speech such as LPC and cepstral coefficients. However, there are several other important features of the original acoustic waveform that are perceptually very important. Two of these are *voicing* and *pitch*, discussed earlier.

The voicing signal, mentioned in section 3.4 above, represents the state of the vibration of the vocal chords. Generally the voicing signal indicates one of three modes. During silences the vocal chords are not vibrating and there is too little airflow through the mouth to create any turbulent sources at constriction points. The result is that little or no sound is produced. During *unvoiced speech* the vocal chords are not vibrating but airflow through constrictions at the tongue, lips or teeth creates turbulent sources which excite the remainder of the vocal tract. The result is breathy or fricative sound. Finally, in *voiced speech* the chords are oscillating producing a approximately periodic impulsive excitation that is subsequently "coloured" or shaped in frequency response by the vocal tract. Voicing can provide imporant perceptual clues about the identity of words and sounds, for example many consonants are distinguished mainly by their *voice onset time.*

Another acoustic feature of interest is *pitch.* Pitch is the term for the rate of excitation of the vocal cords. voiced speech. It is the fundamental frequency of voiced sounds and is thus sometimes known as F0. Tracking pitch can be important for improved signal analysis; for example it is easy to confuse peaks due to resolved pitch harmonics with formant peaks especially for female speakers. In English, pitch also carries information

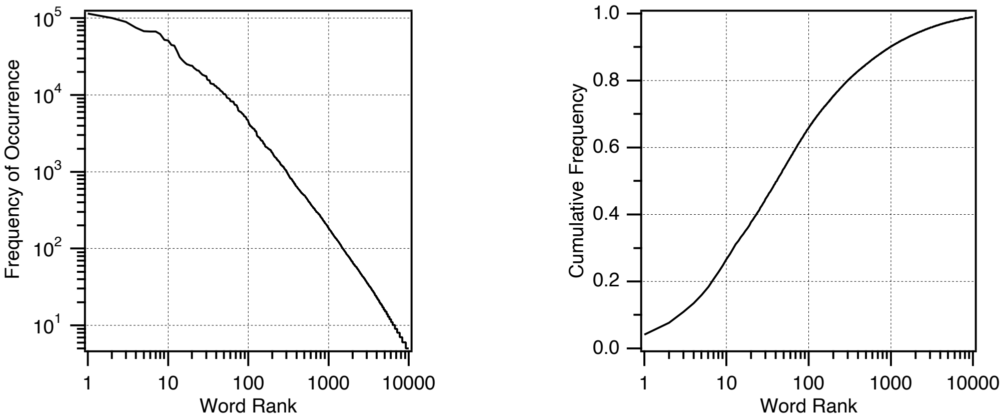

Figure 8: The frequency of common words in spoken language. The left panel shows the frequency of word occurrence compared with word rank indicating that less common words are logarithmically less likely than the more common words in accordance with *Zipf's Law*. The right panel shows the cumulative frequency compared with word rank indicating that even in conversational English roughly 1000 words cover in excess of 90% of the tokens.

about emphasis, questions and parenthetical statements. In tonal languages pitch contours actually carry information about the identity of words and thus pitch estimation is even more desirable. Traditional short-time spectral features such as cepstral coefficients do not capture this.

Automatic algorithms for estimating pitch and voicing exist, but they are typically heuristic and have performance which varies considerably from database to database. Currently, most systems are unable to make use of such cue. Improved estimation of these and other acoustic cues and effective integration of this information into recognition and compression algorithms is one clear direction of research for improving existing speech processing techniques.

4.4 Extra-acoustic cues: language models, lipreading

Speech processing may also benefit from information not present in the acoustic signal. One such piece of information comes from *language models* which attempt to harness the strong statistical patterns in word co-occurrences. The most basic language models measure the unigram, bigram or trigram frequency of words, word doublets or word triplets from enormous spoken text corpora and use these frequencies to evaluate the likelihood of upcoming words in a recognition task. (This is generally difficult to do because of the so-called *Zipf's Law* property of English which means that the most common words occur much more frequently than the less common words. Figure 8 shows the frequency ranking and cumulative frequency of the most common words in conversational speech as measured on the Switchboard database.[6])

Another non-acoustic source of information from which speech processing may also benefit is visual input. Human lipreaders have the ability to infer much of what is

[6] Thanks to Steve Greenberg for these figures.

being said from lip movements alone. But even normal listeners rely heavily on lip information in noisy environments. As video cameras on workstations become more and more common, speech engineers are looking for ways to exploit this new modality to do audio-visual recognition.

5 State of the art: How well are we doing?

State of the art performance for speech recognition still lags far behind that of humans for all tasks from conversational speech transcription to nonsense syllable identification tasks. The systems are also extremely brittle – any small deviation from "normal" conditions causes system performance to degrade dramatically. Even changes that are virtually undetectable to human listeners such as switching the microphone or changing the sampling rate of the audio signal will cause several-fold increases in the error rates of most systems. Furthermore, humans maintain high performance even in the presence of considerable noise while recognition systems again suffer considerably.

To quantify system performance it is common to quote a statistic known as the *word error rate* which is the percentage of words in the reference (true) transcript that the artificial system incorrectly reported. Insertions, substitutions and deletions of words are all counted as errors (though usually tabulated separately). Competitions on standard databases are administered by the National Institute of Standards and Technology (NIST) on a roughly biannual basis. The current evaluation, called Hub4e, uses two databases which contain conversations between two parties over the telephone. These are challenging tasks because they include all of the disfluencies (such as filler words and restarts) that occur in spontaneous speech. They also involve two unknown speakers and an unknown telephone channel distortion. The results are generally quite poor, with most systems getting between 45% and 50% of the words wrong. The systems also run several hundred times slower than real-time on extremely powerful computers. In comparison, humans who listen to the conversations only a few times (making their speed a few times real-time) obtain error rates of only a few percent on the same databases. Table 5 contains the numerical results of this evaluation on the two conversational large vocabulary databases called Switchboard (SB) and Call Home (CH).

System Name	Word Error Rates (%)		
	Switchboard	Call Home	Average
BBN	35.5	53.7	44.9
Boston University	41.5	58.2	50.1
Carnegie Mellon University (ISL)	35.1	54.4	45.1
Cambridge University – HTK	39.2	57.6	48.7
Dragon Systems	39.9	57.4	48.9
SRI	42.5	57.5	50.2

Table 1: Word error rates for state of the art systems in the most recent NIST competition for large vocabulary continuous speech recognition on two conversational telephone speech databases. Results are from the NIST web site [27].

For speech synthesis it is difficult to quantify performance of a system in any way other than having a large group of listeners score generated utterances according to their "naturalness."[7] Thus, by way of example of state of the art synthesis, I have collected generated files from several major synthesizers all saying the same sentence. The sentence was chosen to be one which could not contain pre-memorized phrases, since many synthesis systems have a library of pre-synthesized common phrases which they insert into longer synthesized sentences. The sentence reads: *On Thursdays I polish my vulcanized uncle.*

Such example results are indicative of a general trend in speech synthesizers to be intelligible but not at all natural sounding. In general, speech recognition and speech synthesis are still far behind human performance. Since we use these skills daily, our expectations for artificial systems are quite high.

Speaker identification and speech compression on the other hand are quite advanced. State-of-the-art speaker identification systems compare well to human performance. With hundreds to thousands of users, they can operate at a false rejection rate of 1% and a false acceptance rate of 0.1%. They work for both males and females in the presence of limited noise and do not require a specific phrase to be repeated for identification. Speech compression systems have advanced to the point where *transparent coding* is the de-facto standard. This means that on average listeners cannot tell the difference between the coded version and the original version of an utterance. For speech-only compression (as opposed to general audio including music, etc.) transparent coding can be achieved with a rate as low as 2 kilobits per second, 32 times more compact than a standard telephone channel.

5.1 Overview of current state of the art recognition system architecture

All modern speech recognition systems are based on statistical pattern recognition techniques. The systems are trained on a large number of examples of speech that have been correctly labelled by human transcription. During training, internal parameters of probabilistic models are tuned by algorithms that attempt to maximize how well the models fit the training data. Systems can then be evaluated by showing them new utterances not part of the corpus used to fit parameters and measuring their recognition accuracy. Often part of the training data is reserved for pre-testing diagnosis of the system—known as *validation*. This helps to avoid an *over-fitting* effect in which the models memorize the patterns of training data but do not generalize well to new cases.

Modern system architectures are quite complex and it is certainly beyond the scope of this introduction to describe them in detail. Figure 9 shows a schematic of the basic structure shared by most modern speech recognition systems. An extremely brief description is included, but the reader is refered to the texts [41, 14] for extremely complete reviews. A language model examines the current context of words and proposes several choices for the next word based on co-occurrence statistics. Various pronounciations of these words are looked up in a phonetic dictionary producing candidate phoneme strings.

[7] Although there is much current interest in designing automatic algorithms to predict the results of such human evaluations on synthetic speech (i.e. automatic scoring of synthesis), there is currently no reliable standard.

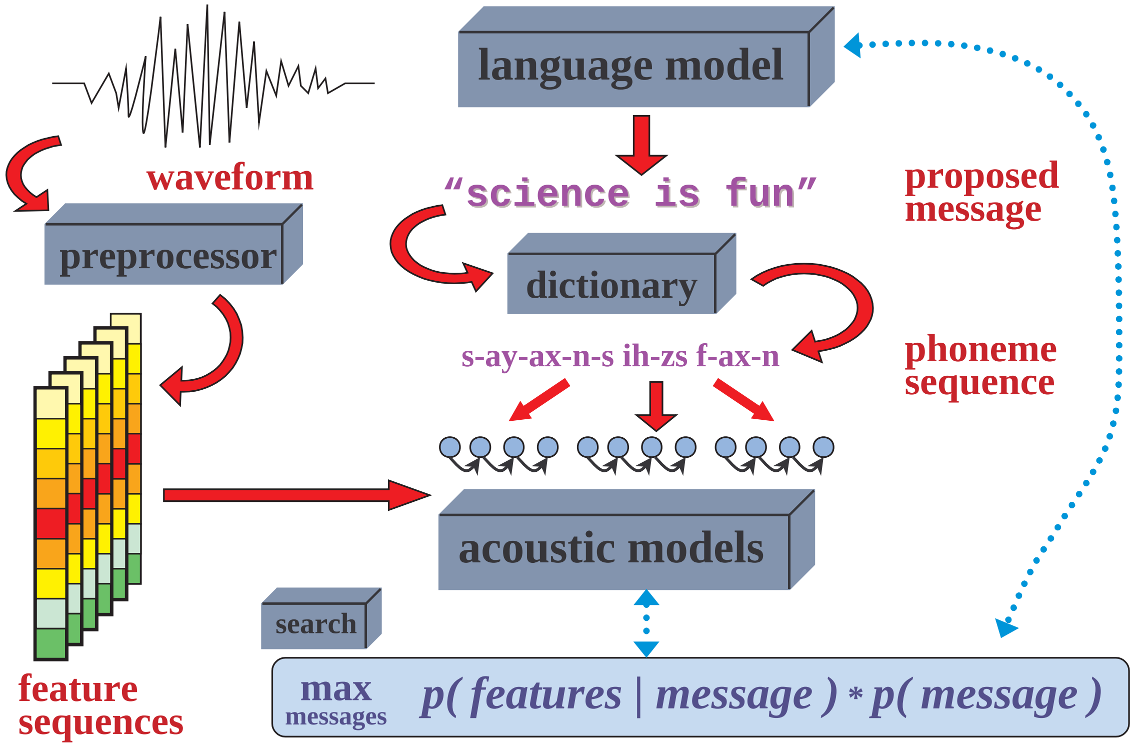

Figure 9: A schematic of the basic structure shared by most modern speech recognition systems. A language model examines the current context of words and proposes several choices for the next word based on co-occurrence statistics. Various pronounciations of these words are looked up in a phonetic dictionary producing candidate phoneme strings. Each phoneme string generates a hidden Markov model composed of the concatenation of the appropriate basic models (e.g. phones or triphones). Each of these generated HMMs gives a probability score to the observed sequence of acoustic features from the incoming speech. This score is combined with a score from the language model indicating the probability of the corresponding word sequence. The choice with the highest combined acoustic and language model score is selected as the result of the recognition process.

Each phoneme string generates a hidden Markov model composed of the concatenation of the appropriate basic models (e.g. phones or triphones). Each of these generated HMMs gives a probability score to the observed sequence of acoustic features from the incoming speech. This score is combined with a score from the language model indicating the probability of the corresponding word sequence. The choice with the highest combined acoustic and language model score is selected as the result of the recognition process.

References

[1] B.S. Atal and S.L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50:637–655, 1972.

[2] L. E. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, 3:1–8, 1972.

[3] L. E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of American Mathematical Society*, 73:360–363, 1967.

[4] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37:1554–1563, 1966.

[5] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[6] R. H. Baumann, J. C. R. Licklider, and B. Howland. Electronic word recognizer. *Journal of the Acoustical Society of America*, 26:137(A), 1954.

[7] A. G. Bell. Patent no. 174,465. U.S. Patent Office, February 14 1876.

[8] A. G. Bell. Prehistoric telephone days. *National Geographic Magazine*, 41:223–242, 1922.

[9] R. Bellman. *Dynamic programming.* Princeton University Press, Princeton, New Jersey, 1957.

[10] H. Bourlard and N. Morgan. *Connectionist Speech Recognition – A Hybrid Approach.* Kluwer Academic Publishers, 1994.

[11] United States Congress. Communications act of 1934. Technical Report 47 USCA – 151, Federal Government, 1934.

[12] E. David and O. Selfridge. Eyes and ears for computers. *Proceedings of the IRE*, pages 1093–1101, May 1962.

[13] H. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits. *Journal of the Acoustical Society of America*, 24(6):637–642, November 1952.

[14] J. Deller, J. Proakis, and J. Hansen. *Discrete Time Processing of Speech Signals.* Macmillan, 1993.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society series B*, 39:1–38, 1977.

[16] P. Denes and M. V. Mathews. Spoken digit recognition using time-frequency pattern matching. *Journal of the Acoustical Society of America*, 32:1450–1455, 1960.

[17] P. B. Denes. On the statistics of spoken English. *Journal of the Acoustical Society of America*, 35(6):892–905, 1963.

[18] Peter Denes and Elliot Pinson. *The Speech Chain: The Physics and Biology of Spoken Language.* W. H. Freeman and Company, New York, 2 edition, 1993.

[19] H. Dudley and S. Balashek. Automatic recognition of phonetic patterns in speech. *Journal of the Acoustical Society of America*, 30:721–732, 1958.

[20] H. Dudley and T. H. Tarnoczy. The speaking machine of wilhelm von kempelen. *Journal of the Acoustical Society of America*, 22, 1950.

[21] Gunnar Fant. *Acoustic Theory of Speech Production.* Mouton and Company, The Hague, 1970 (first published 1960).

[22] Gunnar Fant. *Speech Sounds and Features.* MIT Press, Cambridge, Mass, 1973.

[23] J.D. Ferguson, editor. *Proceedings of the Symposium on the Application of Hidden Markov Models to Text and Speech*, Princeton, NJ, 1980. IDA–CRD.

[24] J. L. Flanagan. *Speech analysis, synthesis, and perception.* Springer-Verlag, New York, 1972.

[25] H. Fletcher. *Speech and hearing in communication.* Van Nostrand, Princeton, NJ, 1953.

[26] H. Fletcher. *Speech and hearing in communication (Van Nostrand 1953), edited by Jont B. Allen.* Acoustical Society of America, New York, NY, ASA edition, 1995.

[27] NIST March 1997 Hub-5E Benchmark Test Results for Recognition of Conversational Speech over the Telephone in English. `ftp://jaguar.ncsl.nist.gov/lvcsr/... mar97/eval/lvcsr_mar97_scores.970410/Summary`. Government web site., April 4 1997.

[28] D. B. Fry and P. Denes. The solution of some fundamental problems in mechanical speech recognition. *Language and Speech*, 1:35–58, 1958.

[29] Dennis Fry. *The Physics of Speech.* Cambridge University Press, Cambridge, 1979.

[30] H. Hassenein and M. Rudko. On the use of discrete cosine transform in cepstral analysis. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-32(4):922, 1984.

[31] A. Holbrook and G. Fairbanks. Dipthong formants and their movements. *Journal of Speecha and Hearing Research*, 5(1):38–58, March 1962.

[32] R. Koenig, H. K. Dunn, and L. Y. Lacy. The sound spectograph. *Journal of Acoustic Society of America*, 18:19–49, 1946.

[33] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4), April 75.

[34] J. D. Markel and Jr. A. H. Gray. *Linear Prediction of Speech.* Springer-Verlag, Berlin, 1976.

[35] A. A. Markov. An example of statistical investigation in the text of 'eugene onyegin' illustrating coupling of 'tests' in chains. In *Proceedings of the Academy of Science, St. Petersburg*, volume 7, pages 153–162, 1913.

[36] Harry F. Olson. *Music, Physics and Engineering.* Dover, 1967.

[37] Douglas O'Shaughnessy. *Speech Communication: Human and Machine.* Addison-Wesley, Reading, 1987.

[38] G.E. Peterson and H.L. Barney. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24:175–184, March 1952.

[39] J. Pierce. Whither speech recognition? *Journal of the Acoustical Society of America*, 46:1049–1051, 1969.

[40] Ralph Potter, George Kopp, and Harriet Kopp. *Visible Speech.* Dover Publications, New York, 1966.

[41] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[42] L. Rabiner and R. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, 1978.

[43] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics Speech and Signal Processing*, 26:43–49, February 1978.

[44] C. P. Smith. A phoneme detector. *Journal of the Acoustical Society of America*, 23:446–451, 1951.

[45] S.S. Stevens and J. Volkman. The relation of pitch to frequency. *American Journal of Psychology*, 53:329, 1940.

[46] T. K. Vintsyuk. Speech discrimination by dynamic programming. *Kibernetka (Cybernetics)*, 4:81–88, January-February 1968.

[47] T. K. Vintsyuk. Element-wise recognition of continuous speech consisting of words from a specified vocabulary. *Kibernetka (Cybernetics)*, pages 133–143, 1971.

[48] Alex Waibel and Kai-Fu Lee. *Readings in Speech Recognition.* Morgan Kaufmann, San Mateo, California, 1990.