LA-UR 96-3519

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE: **AN ARTICULATORILY CONSTRAINED, MAXIMUM ENTROPY APPROACH TO SPEECH RECOGNITION AND SPEECH CODING**

AUTHOR(S): John Hogden

SUBMITTED TO:

External Distribution -Hard Copy

Los Alamos

Los Alamos National Laboratory
Los Alamos New Mexico 87545

DISCLAIMER

An articulatorily constrained, maximum entropy approach to speech recognition and speech coding

John Hogden
CIC-3, MS B265
Los Alamos National Laboratory
Los Alamos, NM 87545

I. Introduction

Hidden Markov models (HMM's) are among the most popular tools for performing computer speech recognition (see Huang, Ariki & Jack, 1990). One of the primary reasons that HMM's typically outperform other speech recognition techniques is that the parameters used for recognition are determined by the data, not by preconceived notions of what the parameters should be. This makes HMM's better able to deal with intra- and inter- speaker variability despite our limited knowledge of how speech signals vary and despite our often limited ability to correctly formulate rules describing variability and invariance in speech. In fact, it is often the case that when HMM parameter values are constrained using our limited knowledge of speech, recognition performance decreases.

However, the structure of an HMM has little in common with the mechanisms underlying speech production. Below, we argue that by using probabilistic models that more accurately embody the process of speech production, we can create models that have all the advantages of HMM's, but that should more accurately capture the statistical properties of real speech samples -- presumably leading to more accurate speech recognition. The model we will discuss uses the fact that speech articulators (the tongue, jaw, lips, etc.) move smoothly and continuously (the word "continuously" is used in the mathematical sense: articulators don't move from one location to another without occupying intermediate positions). Before discussing how to use articulatory constraints, we will give a brief description of HMM's. This will allow us to highlight the similarities and differences between HMM's and the proposed technique.

In a straightforward implementation of the HMM approach, models are made of each word in the vocabulary. The word models are constructed such that we can determine the probability that any speech sample would be produced given a particular word model. The word model most likely to have created a speech sample is taken to be the model of the word that was actually spoken. For example, suppose we produce some new speech sample, **Y**. If w_i is the model for word i, and w_i maximizes the probability of **Y** givenw_i, then a HMM speech recognition algorithm would take word i to be the word which was spoken. In other variants of HMM speech recognition, models are made of phonemes, syllables, or other subword units, and the subword units are recognized.

Figure 1 shows a 5 state HMM applicable to speech recognition (Rabiner & Juang, 1986). Each of the circles in Figure 1 represents a HMM state. At any time, the HMM has one active state and a sound is assumed to be emitted when the state becomes active. The probability of sound y being emitted by state s_i is determined by some parameterized distribution associated with state s_i (e.g. a multivariate Gaussian parameterized by a mean and a covariance matrix). The connections between the states represent the possible interstate transitions. For example, in the Bakis model below, if the model is in state s_2 at time t, then the probability of moving to state s_4 at time $t+1$ is a_{24}

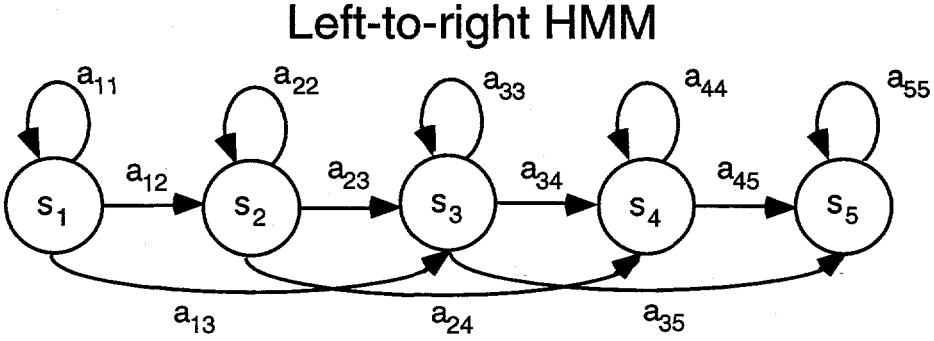

Figure 1

HMM's are trained using a labeled speech data base. For example, the data set may contain several samples of speakers producing the word "president". Using this data, the parameters of the "president" word model (the transition probabilities and the state output probabilities) are adjusted to maximize the likelihood that the "president" word model will output the known speech samples. Similarly, the parameters of the other word models are also adjusted to maximize the likelihood of the appropriate speech samples given the models. We expect that as the word models more closely match the distributions of actual speech samples (i.e. the probability of the data given the word models increases), the recognition performance will improve -- which is why the models are trained in the first place.

One way to make the word models give better estimates of the distributions of speech data is to base the models on the actual processes underlying speech production. Consider that speech sounds are produced by slowly moving articulators (articulator motions have almost all of their energy below 15 Hz. compared to the ~10 kHz acoustic signal). Thus, if we knew the relationship between articulator positions and speech acoustics, we should be able to use information about the articulator positions preceding time t to accurately predict the articulator positions at time t, and therefore predict the acoustic signal at time t. In the following discussion, we show how information about articulation can be used without requiring any training sets other than what is already used to train HMM's. Thus, there is no need to collect extra data about articulator positions, or to use computer simulations to estimate the maping from articulator positions to acoustics.

II. The Model

As with HMM's, in order to determine which sequence of words was most likely to have created the observed data, we want to be able to determine the probability of the observed data given a word model. In the articulatory recognition algorithm presented here, each word will be described in terms of the sequence of articulator positions used to create the word. This is not sufficient, however -- we will also use a parameterized model of the mapping from articulator positions to VQ codes. Using the articulator sequences together with the mapping from articulation to VQ codes gives us a way to estimate the probability of an observed data sequence given a word. In section II.A we describe how an articulator path (a word model) that maximizes the probability of the data can be found, but we will assume that we know about the mapping from speech sounds to articulator positions. In the section II.B we show how to find the mapping from speech sounds to articulator positions using only acoustic speech samples.

II.A Finding Articulatory Trajectories that Maximize the Probability of the Observed Data.

In order to describe how to use articulatorily constrained probabilistic model to perform speech recognition, we will start with some definitions. Let:

n = the number of vector quantization codes in a given speech sample,

$c(t)$ = the VQ code assigned to the t^{th} window of speech,

$\mathbf{c} = [c(1), c(2), \ldots c(n)]$ = a sequence of VQ codes used to describe a speech sample,

$x_i(t)$ = the position of articulator i at time t,

$\mathbf{x}(t) = [x_1(t), x_2(t), \ldots x_d(t)]$ = a vector composed of the positions of all the articulators at time t, and

$\mathbf{X} = [\mathbf{x}(1), \mathbf{x}(2), \ldots \mathbf{x}(n)]$ = a sequence of articulator configurations.

Further definitions are needed to specify the mapping from articulation to VQ codes. Let:

$P(c_i)$ = the probability of observing code c_i given no information about context,

$P(\mathbf{x}|c_i,\varphi)$ = the probability that articulator position $\mathbf{x}$ was used to produce VQ code c_i where

φ = a set of model parameters, e.g. φ could include the mean and covariance matrix of a Gaussian probability density function used to model the distribution of $\mathbf{x}$ given c.

Note that we have left the distributions that give $P(\mathbf{x}|c_i,\varphi)$ unspecified. We have done this because we want to allow for the various possible mappings from acoustics to articulator positions. For example, it has often been argued that many different articulator positions can be used to produce the same acoustic signal (Atal, Chang, Mathews & Tukey, 1978; Schroeter & Sondhi, 1994), although human experiments have not yet verified that this a problem (Hogden et al., 1996; Ladefoged, Harshman, Goldstein & Rice, 1978; Papcun et al., 1992). If there are multimodal distributions of articulator positions that can be used to produce identical acoustic signals, then it may be necessary to specify $P(\mathbf{x}|c_i,\varphi)$ as a mixture of Gaussians.

With these definitions, the probability of observing code c_j given that the current articulator position is $\mathbf{x}$, is expressed as:

$$P(c_j|\mathbf{x},\varphi)=\frac{P(c_j,\mathbf{x}|\varphi)}{P(\mathbf{x}|\varphi)}=\frac{P(c_j,\mathbf{x}|\varphi)}{\sum_i P(c_i,\mathbf{x}|\varphi)}=\frac{P(\mathbf{x}|c_j,\varphi)P(c_j)}{\sum_i P(\mathbf{x}|c_i,\varphi)P(c_i)}$$

Assuming conditional independence, i.e. that $P[c(t)|\mathbf{x}(t),\varphi]$ is independent of $P[c(t')|\mathbf{x}(t'),\varphi]$ for $t\neq t'$:

$$P[\mathbf{c}|\mathbf{X},\varphi]=\prod_{t=0}^{n} P[c(t)|\mathbf{x}(t),\varphi]$$

Note that the probability of observing a code is not independent of the preceding and subsequent codes, it is only conditionally independent. So if $\mathbf{x}(t)$ is dependent on $\mathbf{x}(t')$ then $c(t)$ is dependent on $c(t')$. As demonstrated below, by using an appropriately constrained model of possible articulator trajectories the sequences of codes can be tightly constrained in a biologically plausible manner.

It is possible to find the articulator path that maximizes the probability of a sequence of codes, i.e. find the $\mathbf{X}$ that maximizes $P[\mathbf{c}\,|\,\mathbf{X},\varphi]$, or equivalently, that maximizes $LogP[\mathbf{c}\,|\,\mathbf{X},\varphi]$, where:

$$LogP[\mathbf{c}|\mathbf{X},\varphi]=\sum_t LogP[c(t)|\mathbf{x}(t),\varphi]$$

To show how, we first find $LogP[c(t)|\mathbf{x}(t),\varphi]$:

$$LogP[c(t)|\mathbf{x}(t),\varphi] = Log\frac{P[\mathbf{x}(t)|c(t),\varphi]P[c(t)]}{\sum_i P[\mathbf{x}(t)|c_i,\varphi]P[c_i]}$$

$$= Log\left(P[\mathbf{x}(t)|c(t),\varphi]P[c(t)]\right) - Log\left(\sum_i P[\mathbf{x}(t)|c_i,\varphi]P[c_i]\right)$$

$$= LogP[\mathbf{x}(t)|c(t),\varphi] + LogP[c(t)] - Log\sum_i P[\mathbf{x}(t)|c_i,\varphi]P[c_i]$$

From this, we get:

$$LogP[\mathbf{c}|\mathbf{X},\varphi] = \sum_t LogP[c(t)|\mathbf{x}(t),\varphi]$$

$$= \sum_t \left\{ LogP[\mathbf{x}(t)|c(t),\varphi] + LogP[c(t)] - Log\sum_i P[\mathbf{x}(t)|c_i,\varphi]P[c_i] \right\}$$

Using ∇ to denote the gradient with respect to the components of $\mathbf{x}(t')$, $LogP[\mathbf{c}|\mathbf{X},\varphi]$ is maximized when:

$$\nabla LogP[\mathbf{c}|\mathbf{X},\varphi] = 0 \quad \forall t'$$

Substituting for the left hand side and reducing gives:

$$\nabla\sum_t \left\{ LogP[\mathbf{x}(t)|c(t),\varphi] + LogP[c(t)] - Log\sum_i P[\mathbf{x}(t)|c_i]P[c_i] \right\} = 0 \quad \forall t'$$

$$\sum_t \left\{ \frac{\nabla P[\mathbf{x}(t)|c(t),\varphi]}{P[\mathbf{x}(t)|c(t),\varphi]} + \frac{\nabla P[c(t)]}{P[c(t)]} - \frac{\sum_i \nabla P[\mathbf{x}(t)|c_i,\varphi]P[c_i]}{\sum_i P[\mathbf{x}(t)|c_i,\varphi]P[c_i]} \right\} = 0 \quad \forall t'$$

$$\frac{\nabla P[\mathbf{x}(t')|c(t'),\varphi]}{P[\mathbf{x}(t')|c(t'),\varphi]} + \frac{\nabla P[c(t')]}{P[c(t')]} - \frac{\nabla\sum_i P[\mathbf{x}(t')|c_i,\varphi]P[c_i]}{\sum_i P[\mathbf{x}(t')|c_i,\varphi]P[c_i]} = 0 \quad \forall t'$$

concluding with the equation:

$$\nabla LogP[\mathbf{c}|\mathbf{X},\varphi] = \frac{\nabla P[\mathbf{x}(t')|c(t'),\varphi]}{P[\mathbf{x}(t')|c(t'),\varphi]} - \frac{\sum_i P[c_i]\nabla P[\mathbf{x}(t')|c_i,\varphi]}{\sum_i P[\mathbf{x}(t')|c_i,\varphi]P[c_i]} = 0 \quad \forall t'$$

The preceding analysis is incomplete because it ignores constraints on the possible articulator paths. To incorporate biologically plausible constraints on articulator motion, we will allow only those articulator trajectories that have all their energy below some cut-off frequency (say 15 Hz, since actual articulator paths have very little energy above 15 Hz). The constraint that the articulator path have all of its energy below the cut-off frequency is equivalent to requiring that the path lie on a hyperplane composed of the axes defined by low frequency sine and cosine waves.

When $\nabla Log(\mathbf{c}|\mathbf{X},\varphi)$ is perpendicular to the constraining hyperplane, so that $Log(\mathbf{c}|\mathbf{X},\varphi)$ can not increase without traveling off of the hyperplane, then we have reached a constrained local minimum. Thus, the smooth path that maximizes the likelihood of the observed data is the path for which $\nabla Log(\mathbf{c}|\mathbf{X},\varphi)$ has no components with energy below the cut-off frequency. This suggests the following algorithm for finding the smooth path that maximizes the probability of the data:

1) start with any smooth path.

2) find the gradient of the log probability of the smooth path.

3) low-pass filtered the gradient to determine the gradient projected onto the constraining hyperplane.

4) add the low-pass filtered gradient times some small constant to the path to get a better estimate of the most likely smooth path.

5) repeat steps 2 - 4 until the algorithm converges.

There are also a variety of standard numerical algorithms that can be used to maximize functions. Using one of these algorithms can speed up the process of finding the most likely smooth path. One of these techniques, the conjugate gradient algorithm, is used in the current implementation.

One additional point should be made here: the path which maximizes the conditional probability of the data is also the path that minimizes the number of bits that need to be transmitted in addition to the smooth path to specify the data. This can be seen from information theory (Sayood, 1996), which shows that the number of bits that must be transmitted in addition to the smooth path is:

$$\text{bits} = \sum_t \left(1 - LogP\left[c(t)|\mathbf{x}(t),\varphi\right]\right) = \sum_t 1 - \sum_t LogP\left[c(t)|\mathbf{x}(t),\varphi\right]$$

Since we are maximizing $\sum_t LogP\left[c(t)|\mathbf{x}(t),\varphi\right]$, we are minimizing the number of bits. This result suggests that this approach has potential as a speech coding technique.

II.B Finding a Mapping from Articulation to Acoustics

In the preceding section, we assumed that we knew $P(c)$ and $P(\mathbf{x}|c,\varphi)$. In this section we show that these values can be determined using only acoustic data. This is an important section, because $P(\mathbf{x}|c,\varphi)$ is a probabilistic mapping from speech sounds to articulator positions, and our claim is that this mapping can be inferred using only acoustic data. To emphasize the importance of this section, consider what would previously have been required to calculate $P(\mathbf{x}|c,\varphi)$. If we had a sufficiently large speech database containing both acoustics and measurements of articulator positions, then it would not be difficult to calculate $P(\mathbf{x}|c,\varphi)$ -- we would merely need to perform VQ on all the acoustic signals and then find the distribution of $\mathbf{x}$ for each code (Hogden et al., 1996). Alternately, we could use a neural network or some other form of nonlinear regression to find the relationship between acoustics and articulation (Ladefoged et al., 1978; Papcun et al., 1992; Zlokarnik, 1995). Unfortunately, collecting a sufficiently large database of articulator measurements, including information about the velum, pharynx, lips, tongue, and jaw would be impractical. To avoid the difficulty of collecting inordinate amounts of articulator position measurements, we could use a model of the vocal tract to generate the speech sounds from known (albeit modeled) articulator positions (Atal et al., 1978; Boe, Perrier & Bailly, 1992; Rahim, Kleijn, Schroeter &

Goodyear, 1991; Schroeter & Sondhi, 1992; Schroeter & Sondhi, 1994). This approach is also problematic because there is reason to believe that the relationship between acoustics and articulation for human speech may be very different than for articulatory speech synthesizers, and, in fact, may be very different with different articulatory speech synthesizers (Hogden et al., 1996).

However, using maximum likelihood estimation, it is possible to find a good approximation of the relationship between acoustics and articulations by building on the framework presented above. All we have to do is iteratively repeat two steps:

1) given a collection of quantized speech signals and some initial estimate of the mapping from acoustics to speech, use the procedures in the preceding section to find the paths that maximize the probability of the observed data.

2) given the paths that maximize the probability of the data, find the value of φ and the $P(c_i)$ values that will increase the probability of the data.

Since both of these steps will increase the probability of the data, by iteratively repeating them, we will increase the probability of the data until we have reached a local (possibly global) maximum.

It is easy to find the maximum likelihood estimate of $P(c)$ given enough speech samples -- the model $P(c)$ values should be set equal to the observed probabilities of c. Calculating φ can be accomplished using standard maximization algorithms. Maximization algorithms that use gradient information are typically faster than algorithms that don't use the gradient, making it advantageous to have an expression for $\nabla \mathrm{Log}(\mathbf{c}|\mathbf{X},\varphi)$ with respect to φ. This expression can be derived as below:

$$\nabla LogP[\mathbf{c}|\mathbf{X},\varphi]$$

$$= \nabla \sum_t \left\{ LogP[\mathbf{x}(t)|c(t),\varphi] + LogP[c(t)] - Log \sum_i P[\mathbf{x}(t)|c_i]P[c_i] \right\}$$

$$= \sum_t \left\{ \frac{\nabla P[\mathbf{x}(t)|c(t),\varphi]}{P[\mathbf{x}(t)|c(t),\varphi]} + \frac{\nabla P[c(t)]}{P[c(t)]} - \frac{\sum_i \nabla P[\mathbf{x}(t)|c_i,\varphi]P[c_i]}{\sum_i P[\mathbf{x}(t)|c_i,\varphi]P[c_i]} \right\}$$

$$= \sum_t \left\{ \frac{\nabla P[\mathbf{x}(t)|c(t),\varphi]}{P[\mathbf{x}(t)|c(t),\varphi]} - \frac{\sum_i \nabla P[\mathbf{x}(t)|c_i,\varphi]P[c_i]}{\sum_i P[\mathbf{x}(t)|c_i,\varphi]P[c_i]} \right\}$$

Thus, using only acoustic speech samples, it is possible to derive the relationship between acoustics and articulation, and also the articulator paths that are most likely to have created a speech sample.

References

Atal, B. S., Chang, J. J., Mathews, M. V., & Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. Journal of the Acoustical Society of America, 63(5), 1535-1555.

Boe, L. J., Perrier, P., & Bailly, G. (1992). The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory conversion. Journal of Phonetics, 20, 27-38.

Hogden, J., Zlokarnik, I., Lofqvist, A., Gracco, V., Rubin, P., & Saltzman, E. (1996). Accurate recovery of articulator positions from acoustics -- new conclusions based on human data. Journal of the Acoustical Society of America, 100(3).

Huang, X. D., Ariki, Y., & Jack, M. (1990). Hidden Markov Models for Speech Recognition. Edinburgh: Edinburgh University Press.

Ladefoged, P., Harshman, R., Goldstein, L., & Rice, L. (1978). Generating vocal tract shapes from formant frequencies. Journal of the Acoustical Society of America, 64(4), 1027-1035.

Papcun, G., Hotchberg, J., Thomas, T., Laroche, F., Zacks, J., & Levy, S. (1992). Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data. Journal of the Acoustical Society of America, 92(2), 688-700.

Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. IEEE ASSP Magazine.

Rahim, M. G., Kleijn, W. B., Schroeter, J., & Goodyear, C. C. (1991). Acoustic to articulatory parameter mapping using an assembly of neural networks. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 485-488.

Sayood, K. (1996). Introduction to Data Compression. San Francisco, CA: Morgan Kaufmann Publishers, Inc.

Schroeter, J., & Sondhi, M. (1992). Speech coding based on physiological models of speech production. In S. Furui & M. Sondhi (Eds.), Advances in Speech Signal Processing, (pp. 231-267). New York: Marcel Dekker, Inc.

Schroeter, J., & Sondhi, M. (1994). Techniques for estimating vocal-tract shapes from the speech signal. IEEE Transactions on Speech and Audio Processing, 2(1), 133-150.

Zlokarnik, I. (1995). Adding articulatory features to acoustic features for automatic speech recognition. Journal of the Acoustical Society of America, 97(5 pt. 2), 3246(A).