LA-UR- 98-1378

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: **MALCOM X: COMBINING MAXIMUM LIKELIHOOD CONTINUITY MAPPING WITH GAUSSIAN MIXTURE MODELS**

AUTHOR(S): John Hogden, James C. Scovel

SUBMITTED TO: External Distribution - Hard Copy

Los Alamos

Los Alamos National Laboratory
Los Alamos New Mexico 87545

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

MALCOM X: combining maximum likelihood continuity mapping with Gaussian mixture models

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

A Gaussian mixture model (GMM) of a digitized speech waveform can be created by 1) breaking the waveform into successive windows of speech, 2) distilling each window of speech into a vector of acoustic features, $\mathbf{a}(t)$, where t indicates the temporal position of the window, and 3) finding a set of component Gaussian distributions indexed by c, $P(\mathbf{a}|c)$, and a set of a priori probabilities, $P(c)$, that maximize the probability of the speech data, i.e. that maximize

$$P(\mathbf{A} \mid \lambda) = \prod_t \sum_c P(c) P[\mathbf{a}(t)|c] \qquad \text{EQ. 1}$$

where $\mathbf{A}$ represents the sequence of $\mathbf{a}(t)$ values over all t and λ comprises the GMM model parameters (Reynolds & Rose, 1995, gives a detailed description of GMMs). GMMs are among the best speaker recognition algorithms currently available. However, the GMM's estimate of the probability of the speech signal does not change if we randomly shuffle the temporal order of the feature vectors, even though the actual probability of observing the shuffled signal would be dramatically different -- probably near zero. A potential way to improve the performance of GMMs is to incorporate temporal information into the estimate of the probability of the data. Doing so could improve speech recognition, speaker recognition, and potentially aid in detecting lies (abnormalities) in speech data.

As described in other documents (Hogden, 1996), MALCOM is an algorithm that can be used to estimate the probability of a sequence of categorical data. MALCOM can also be applied to speech (and other real valued sequences) if windows of the speech are first categorized using a technique such as vector quantization (Gray, 1984). However, by quantizing the windows of speech, MALCOM ignores information about the within-category differences of the speech windows. Thus, MALCOM and GMMs complement each other: MALCOM is good at using sequence information whereas GMMs capture within-category differences better than the vector quantization typically used by MALCOM.

An extension of MALCOM (MALCOM X) that can be used for estimating the probability of a speech sequence is described below. MALCOM X combines features of a GMM with features of MALCOM. As in MALCOM, in MALCOM X it is assumed that an unobservable object moving smoothly through the abstract space called a *continuity map* (CM) periodically produces a categorical data value, where the probability of producing a particular categorical data value at time t is a function of the position of the unobservable object at time t, $\mathbf{x}(t)$, and the model parameters, φ. However, in MALCOM X, the categorical data value is thought of as the index of a Gaussian distribution giving $P(\mathbf{a}|c)$, where $\mathbf{a}$ and c have the same meaning as in the GMM.

Thus, the probability of observing the acoustic vector $\mathbf{a}(t)$ is:

$$P[\mathbf{a}(t)|\mathbf{x}(t),\varphi] = \sum_c P[\mathbf{a}(t)|c,\mathbf{x}(t),\varphi] P[c|\mathbf{x}(t),\varphi] \qquad \text{EQ. 2}$$

Making the assumption that

$$P[\mathbf{a}(t)|c,\mathbf{x}(t),\varphi] = P[\mathbf{a}(t)|c,\varphi] \qquad \text{EQ. 3}$$

and the typical MALCOM assumption of conditional independence, we can write the probability of the sequence of acoustic vectors as:

$$P(\mathbf{A}|\mathbf{X},\varphi) = \prod_t \sum_c P[\mathbf{a}(t)|c,\varphi]P[c|\mathbf{x}(t),\varphi] \qquad \text{EQ. 4}$$

where $\mathbf{X}$ represents a sequence of $\mathbf{x}(t)$ positions, i.e. a path through the CM. Finally, to get the probability of $\mathbf{A}$ we need to integrate over all smooth $\mathbf{X}$:

$$P(\mathbf{A}|\varphi) = \int P(\mathbf{X}|\varphi)\prod_t \sum_c P[\mathbf{a}(t)|c,\varphi]P[c|\mathbf{x}(t),\varphi]d\mathbf{X} \qquad \text{EQ. 5}$$

where:

$$P(\mathbf{X}|\varphi) = \begin{cases} 0 & \text{if } \mathbf{X} \text{ is not a smooth path} \\ \varepsilon & \text{if } \mathbf{X} \text{ is a smooth path} \end{cases} \qquad \text{EQ. 6}$$

and

$$\int P(\mathbf{X}|\varphi)d\mathbf{X} = 1 \qquad \text{EQ. 7}$$

Thus, it is possible to write an equation giving the probability of the sequence of acoustic vectors given the MALCOM X model. However, calculating the probability would be impractical because of the difficulty of performing the required integration. The difficulties of finding the solution call for a sub-optimal approach. For example, the Viterbi algorithm is often used as a sub-optimal approximation of the probability found using the hidden Markov model (HMM) forward algorithm (Rabiner & Juang, 1986, prompted the following discussion). It is worthwhile to look more closely at the Viterbi algorithm to see the similarity to the current problem. To make the analogy clear, let $\mathbf{X}'$ represent a sequence of HMM states and let φ' be the HMM model parameters. The Viterbi algorithm finds the $\mathbf{X}'$ that maximizes $P(\mathbf{A},\mathbf{X}'|\varphi')$. However, as Rabiner points out, the Viterbi algorithm can be used to approximate $P(\mathbf{A}|\varphi')$. The relationship between $P(\mathbf{A},\mathbf{X}'|\varphi')$ and $P(\mathbf{A}|\varphi')$ is:

$$P(\mathbf{A}|\varphi') = \sum_{\mathbf{X}'} P(\mathbf{A},\mathbf{X}'|\varphi') \qquad \text{EQ. 8}$$

As can be seen from EQ. 8, the Viterbi algorithm calculates only one term in the summation. However, the Viterbi algorithm is a reasonable approximation because it tends to be the case that the term calculated by the Viterbi algorithm is the only significant term in the summation. Interestingly, EQ 5, the equation for the probability of the acoustic sequence given the MALCOM X model, can be rewritten as:

$$P(\mathbf{A}|\varphi) = \int p(\mathbf{A},\mathbf{X}|\varphi)d\mathbf{X} \qquad \text{EQ. 9}$$

which bears a close similarity to EQ. 8. This similarity suggests finding the $\mathbf{X}$ which maximizes EQ. 9, i.e.

$$\hat{\mathbf{X}}(\mathbf{A},\varphi) = \arg\max_{\mathbf{X}} P(\mathbf{A},\mathbf{X}|\varphi) \qquad \text{EQ. 10}$$

and then using the approximation:

$$P(\mathbf{A}|\varphi) \cong P(\mathbf{A},\hat{\mathbf{X}}(\mathbf{A},\varphi)|\varphi)$$
EQ. 11

to get our estimate of the probability of the sequence of acoustic vectors. Clearly, the approximation in EQ. 11 underestimates $P(\mathbf{A}|\varphi)$, since it only includes one term from the integral. However, it is not clear to what extent the inaccuracy will affect speech and/or speaker recognition. In fact, GMMs also must underestimate $P(\mathbf{A}|\varphi)$ for some signals, as implied by the fact GMMs tend to overestimate the probability of signals which are very unlikely -- such as signals made by randomly shuffling the order of the acoustic vectors.

A second difficulty of using EQ. 11 is that EQ. 11 gives the height of a probability density function (PDF) -- not a probability. This is problematic for various reasons, not the least of which is that the height of a PDF does not have to lie between 0 and 1 as a probability would. This is not a particularly serious problem, particularly considering that GMMs also return the heights of PDFs, not probabilities. In fact, the extent to which the approximation in EQ. 11 must hold is determined by the problem to be solved. Consider a typical speaker identification scenario, in which the most probable speaker is chosen using the rule:

$$\hat{i} = \arg\max_i P(\varphi_i|\mathbf{A}) = \arg\max_i \frac{P(\mathbf{A}|\varphi_i)P(\varphi_i)}{P(\mathbf{A})} = \arg\max_i \frac{P(\mathbf{A}|\varphi_i)P(\varphi_i)}{\sum_j P(\mathbf{A}|\varphi_j)P(\varphi_j)} \qquad \text{EQ. 12}$$

where i identifies the speaker so φ_i is the model for speaker i. The first thing to notice about EQ. 12 is that

$$0 \le \frac{P(\mathbf{A}|\varphi_i)P(\varphi_i)}{\sum_j P(\mathbf{A}|\varphi_j)P(\varphi_j)} \le 1 \qquad \text{EQ. 13}$$

i.e., the value has the same range as a probability. Realizing that $P(\mathbf{A})$ is constant, if we treat $P(\varphi_i)$ as constant then the speaker with the highest value of $P(\mathbf{A}|\varphi)$ is the speaker that will be chosen. So for this speaker identification task, we only need to believe that

$$\arg\max_i P(\mathbf{A}|\varphi_i) = \arg\max_i P(\mathbf{A},\hat{\mathbf{X}}(\mathbf{A},\varphi_i)|\varphi_i) \qquad \text{EQ. 14}$$

The same arguments can be made for word recognition, where the word model most likely to have produced the data is chosen.

While the shortcomings of MALCOM X do call for further refinements, one should remember that probability estimates coming from models like GMMs, HMMs and MALCOM X are only accurate to the extent that the models reflect the underlying data production process. Since current models of speech data are rather simplistic, it seems likely that the error due to other model inaccuracies outweigh the error due to the approximation in EQ. 11.

The approximation in EQ. 11 is only useful if there is a practical way to find $\hat{\mathbf{X}}(\mathbf{A},\varphi)$. We can solve for $\hat{\mathbf{X}}(\mathbf{A},\varphi)$ in much the same way as we solve other problems in MALCOM -- by using standard gradient maximization techniques. As in many probability maximization problems, the logarithm of the probability is maximized to simplify the derivation. The gradient of the logarithm of $P(\mathbf{A},\mathbf{X}'|\varphi')$ with respect to $\mathbf{X}$ is:

$$\nabla Log P[\mathbf{A},\mathbf{X}|\varphi] = \nabla P(\mathbf{X}|\varphi)\sum_t Log \sum_c P[\mathbf{a}(t)|c]\nabla P[c|\mathbf{x}(t),\varphi]$$

EQ. 15

$$= \frac{\sum_c P[\mathbf{a}(t)|c]\nabla P[c|\mathbf{x}(t),\varphi]}{\sum_c P[\mathbf{a}(t)|c]P[c|\mathbf{x}(t),\varphi]} \quad \text{for smooth } \mathbf{X}$$

Since $\nabla P[c|\mathbf{x}(t),\varphi]$ has been simplified in previous descriptions of MALCOM, we will not continue the derivation here.

It is also possible to find the maximum likelihood values for the MALCOM X model parameters. A variant of the EM algorithm can be used for this purpose. Essentially, the required steps are 1) given a model and speech data, find the $\mathbf{X}$ which maximizes the probability of the data; 2) change the model parameters to increase the probability of the data given the $\mathbf{X}$ values calculated in step 1. The method for performing step 1 was described above. Step 2 has two subcomponents 1) find the parameters associated with the distributions over continuity map position and 2) find the parameters associated with the distributions over acoustic vectors. These steps are essentially the same as the techniques previously described MALCOM and for GMMs respectively. MALCOM X can also be trivially extended to data sets containing more than one speech sequence.

References

Gray, R. (1984). Vector Quantization. IEEE Acoustics, Speech, and Signal Processing Magazine, 4-29.

Hogden, J. (1996). Improving on hidden Markov models: An articulatorily constrained, maximum likelihood approach to speech recognition and speech coding (unclassified LA-UR-96-3945). Los Alamos, NM: Los Alamos National Laboratory.

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

Reynolds, D., & Rose, R. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing, 3(1), 72-83.