LA-UR- 00-1084

Approved for public release; distribution is unlimited.

Title: BRIDGING THE GAP BETWEEN SPEECH PRODUCTION AND SPEECH RECOGNITION

Author(s): JOHN E. HOGDEN, CIC-3
PATRICK F. VALDEZ, CIC-3

Submitted to: SPEECH PRODUCTION SEMINAR
MAY 1-4, 2000
GERMANY

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

BRIDGING THE GAP BETWEEN SPEECH PRODUCTION AND SPEECH RECOGNITION

John Hogden & Patrick Valdez
Los Alamos National Laboratory
hogden@lanl.gov

ABSTRACT

Although stochastic models of speech signals (e.g. hidden Markov models, trigrams, etc) have lead to impressive improvements in speech recognition accuracy, it has been noted that these models have little relationship to speech production (Lee, 1989) and their recognition performance on some important tasks is far from perfect. However, there have been recent attempts to bridge the gap between speech production and speech recognition using models that are stochastic and yet make more reasonable assumptions about the mechanisms underlying speech production (Bakis, 1991; Deng, 1998; Hogden, 1998; Picone et al., 1999). One of theses models, Multiple Observable, Maximum Likelihood Continuity Mapping (MO-MALCOM) is described in this paper.

There are theoretical and experimental reasons to believe that MO-MALCOM learns an invertable stochastic mapping between articulator positions and speech acoustics. Furthermore, MO-MALCOM can be combined with standard speech recognition algorithms to create a speech recognition model based on a stochastic production model. Results of using MO-MALCOM speech recognition on data derived from the switchboard corpus will be discussed.

BACKGROUND: STATE-OF-THE-ART SPEECH RECOGNITION

In many realistic domains, automatic speech recognition performance (ASR) is inadequate. To be concrete, at the National Institute of Standards and Technology 1998 HUB-5 Speech Recognition Evaluation, state-of-the-art systems had about a 60%-65% word recognition rate on "casual speech", i.e., telephone conversations in the Switchboard database (Martin, Fiscus, Przybocki & Fisher, 1998). Since speaking rates of 200 words per minute are not uncommon in casual speech, a 60% word recognition accuracy implies approximately 80 errors per minute -- an unacceptably high rate for many applications. Furthermore, recognition performance is not improving rapidly. Improvements in word recognition accuracy of a few percent are considered "big" improvements, and recognition rates of the best systems on the Switchboard data were between 64.9% and 61.2% for 1996, 1997, and 1998, although they have improved from only 52% recognition in the 1995 evaluation[1]. These types of recognition results should prompt us to look for alternatives to the current approach.

The primary tool used in speech recognition is the hidden Markov model (HMM) -- they are used to estimate the probability of an acoustic sequence given the model parameters (Jelinek, 1997). A nice feature of HMMs is that maximum likelihood techniques allow the model parameters to be automatically determined from training data. The automatic parameter estimation, and the stochastic nature of the HMMs are presumably the features that allow them to cope with the amazing amount of variability in speech.

While HMMs have been useful, few researchers would argue that speech production is a Markov process, or even a hidden Markov process. In fact, it has been noted that "[the HMM] is a very inaccurate model of the speech production process" (Lee, 1989). The problems with HMMs have prompted many researchers to propose alternatives (a good review is given in Ostendorf, Digalakis & Kimball, 1996). Most of the alternatives add parameters to HMMs to allow greater ability to model signals. Adding parameters has the disadvantage that more data is needed to train the models, and training data sets are already very large. In our opinion, making the acoustic models more general by adding parameters is the wrong way to go. In fact, like other researchers in the field (Bakis, 1991; Deng, 1998; Picone et al., 1999), we are interested in making the acoustic models more specific to speech, i.e., retain the stochastic nature of the model and the automatic parameter estimation, but change the underlying model to more accurately represent speech production.

INTRODUCING MO-MALCOM

Multiple Observable Maximum Likelihood continuity Mapping (MO-MALCOM) is a variant of the MALCOM algorithm. As such, it produces a stochastic model of two or more sequences of categorical data values. In the work reported here, the categorical data sequences are vector quantization (VQ) codes, representing speech acoustics, and time-aligned phoneme labels.

The main assumptions of MO-MALCOM are 1) that data sequences are produced by objects moving smoothly through an abstract space called a continuity map (CM) and 2) that the probability of observing a particular data value at time t is dependent on the position of the object at time t, x(t). These assumptions model the facts that 1) speech sounds are produced by motions of the speech articulators and 2) the sound output at time t can be determined from the positions of the articulators at time t. As discussed below, these assumptions are realistic enough that it is often helpful to think of CM positions as estimated articulator positions.

MO-MALCOM is very similar in spirit to an HMM. As with HMMs, the MO-MALCOM parameters are learned from training data using maximum likelihood techniques. Paths of the objects through the continuity map are considered unobservable and are inferred in much the same way as HMM state sequences. Furthermore, the probability of outputting a

[1] There was no 1999 evaluation and the 2000 evaluation has not started as of the time this paper was written

particular VQ code given a path position is analogous to the output probability of an HMM state. However, when we continue the analogy of MO-MALCOM continuity map position to states we see two major differences between HMMs and MALCOM: 1) in an HMM there are a finite number of states whereas in MALCOM the state is a continuous variable; 2) MALCOM uses a smoothness constraint on paths through the CM, which is much more realistic and uses much more context than a first or second order Markov assumption.

In order to determine the probability of each code for each of an infinite number of CM positions, MO-MALCOM estimates the parameters of probability density functions (e.g. Gaussians) over the continuity map, which quantify the probability of a position in the continuity map given a VQ code. To be concrete, for each VQ code, c_i, MO-MALCOM estimates an a priori probability, $P[c_i]$, and parameters of PDFs that give $P[\mathbf{x}|c_i,\varphi]$, where $\mathbf{X}$ denotes a position in the CM and φ represents the PDF parameters (e.g. means and covariance matrices). These parameters constitute the MO-MALCOM estimate of a probabilistic mapping between VQ codes and articulation.

While it may or may not be possible to invert a deterministic mapping from articulation to acoustics (Atal, Chang, Mathews & Tukey, 1978; Hogden et al., 1993), Bayes' law makes it easy to invert MO-MALCOM's probabilistic mapping to get the probability of a VQ code given a continuity map position.

$$P[c_i|\mathbf{x},\varphi] = \frac{P[\mathbf{x}|c_i,\varphi]P[c_i]}{P[\mathbf{x}|\varphi]}$$

and analogous techniques are use to get the probability of each phone, f_i, given a continuity map position:

$$P[f_i|\mathbf{x},\varphi] = \frac{P[\mathbf{x}|f_i,\varphi]P[f_i]}{P[\mathbf{x}|\varphi]}$$

The joint probability of fi and ci given a CM positions is given by:

$$P[f_i,c_i|\mathbf{x},\varphi] = P[f_i|\mathbf{x},\varphi]P[c_i|\mathbf{x},\varphi]$$

and the joint probability of a sequences of VQ codes, **c** = [c(1), c(2), ... c(n)], and phones, **f** = [f(1), f(2), ... f(n)], given a path through the CM, **X** = [**x**(1), **x**(2), ... **x**(n)], is given by the equation:

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

As with HMMs, we must make a conditional independence assumption to calculate the probability of a whole sequence from the probabilities at each time. However, this assumption is somewhat more warranted for MO-MALCOM. To support this claim, note that if the path position at time t contains sufficient information about the articulator positions at time t, then one should expect conditional independence – just as one should expect that the sound output from the mouth at time t depends only on the articulator positions at time t. As we discuss below, there is good reason to believe that MO-MALCOM is capable of inferring articulator positions, so the conditional independence assumptions for VQ codes is probably not too bad. However, it is unlikely that the probability of a phone at t given the articulator positions at t is conditionally independent of the temporal context, so the MO-MALCOM assumption can likely be improved. Further discussion of this point can be found below.

Signal Processing

Before applying MO-MALCOM to continuous valued data, short time-windows of the data should be processed into vectors that contain information about vocal-tract shape and as little information as possible about the vocal-tract excitation. (e.g. cepstra, LPC coefficients, mel-cepstra). This signal processing must be done to meet the MO-MALCOM assumption that the signals are produced by slowly moving objects, such as the articulators, not quickly moving objects such as the vocal chords. The resulting sequences of vectors are then converted to sequences of categorical data values using VQ.

MO-MALCOM Training

As with HMMs, the MO-MALCOM parameters need to be trained on a large corpus of training data. Two learning steps are iteratively repeated to calculate the parameters of the PDFs:

1) Given some initial set of PDF parameters, and many different examples of simultaneous **f** and **c** data sequences, find the smooth paths (i.e. paths that have no Fourier components above some cut-off frequency) through the CM that maximize the likelihood of the code sequences. That is, representing the path though the CM by **X** = [**x**(1), **x**(2), ... **x**(n)], and the PDF parameters by φ, find

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

2) Find the values of the PDF parameters that maximize the probability of the data category sequence. That is:

$$\hat{\varphi} = \arg\max_{\varphi} P[\mathbf{c},\mathbf{f}|\hat{\mathbf{X}},\varphi]$$

There are a variety standard algorithms for performing the maximizations required above. We have found conjugate gradient ascent methods useful.

MO-MALCOM Speech Recognition

During recognition, we first want to estimate the probability of each phone at each time step. To do so we first find the smooth path throught the CM that maximizes the probability of the VQ codes:

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

After obtaining this path estimate, it is possible to estimate the probability of each phone at each time.

Combining MO-MALCOM with Word Models

In standard speech recognition algorithms, the probability of a phone sequence given a word, $P\left[f(t) = f_i \middle| w\right]$, is estimated using a lattice model, and is then used to get the probability of a word given the observable data. Since a variety of standard techniques can be used to create a lattice model, we will not discuss the problem of estimating lattice model structures or parameters here. However, in the next section, we discuss one way to combine a lattice structure with MO-MALCOM processing to achieve speech recognition.

Define variables reminiscent of the HMM forward algorithm:

$$a_{ij} = P\left[f(t) = f_i \middle| f(t-1) = f_j, w\right]$$

$$b_i(t) = P\left[f(t) = f_i \middle| x(t)\right]$$

$$\pi_i = P\left[f(1) = f_i \middle| w\right]$$

$$\alpha_i(t) = P\left[f(t) = f_j \middle| w, x(t), x(t-1), x(t-2), \ldots, x(1)\right]$$

The reader can confirm that

$$\alpha_i(1) = b_i(1)\pi_i$$

and

$$\alpha_i(t) = b_i(t)\sum_j a_{ij}\alpha_j(t-1)$$

Using these recursively calculated probabilities we can find

$$P\left[w \middle| \mathbf{X}\right] = P\left[w \middle| x(t), x(t-1), x(t-2), \ldots, x(1)\right] = \sum_i \alpha_i(t)P(w)$$

Ideally, we would find

$$P\left[w \middle| \mathbf{c}\right] \cong \int P\left[w \middle| \mathbf{X}\right] P\left[\mathbf{X} \middle| \mathbf{c}\right] d\mathbf{X}$$

but since calculating this integral is not practical, we will assume that $P\left[\mathbf{X} \middle| \mathbf{c}\right]$ is only non-zero in an infinitesimal region around

$$\hat{\mathbf{X}} = \arg\max_{\mathbf{X}} P\left[\mathbf{c} \middle| \mathbf{X}\right]$$

and calculate

$$P\left[w \middle| \mathbf{c}\right] \cong P\left[w \middle| \hat{\mathbf{X}}\right]$$

The basic idea, then, is to start by finding the smooth path through the continuity map that maximizes the probability of the VQ code sequence. Then use that path to get the probability of each phoneme for each acoustic window. Then combine the probabilities of each phoneme given the path, with the phoneme probabilities given the word, and the prior word probability, to get an estimate of the posterior probability of the word.

Optimizing Other MO-MALCOM Parameters

The formulation above assumes that we know the number of dimensions to use in the continuity map, and that we know what cut-off frequency to use to constrain the smooth paths. Of course, these parameters are typically not known a priori. The obvious way to determine the parameters is to simply try many combinations of parameters and determine which combination works best for the problem being studied. However, doing so can be very time consuming.

A (sometimes) more expedient way to optimize the number of dimensions and cut-off frequency is to cross-validate using MO-MALCOM's estimate of the probability of a cross-validation set as a measure of how well the model is performing. Doing so is relatively straightforward, but it should be remembered that estimating MO-MALCOM paths from the cross-validation set, and then determining the probability of the data given the paths, will result in a biased estimate of the generalization performance. Instead, a MO-MALCOM path should be estimated without using one pair of categorical data values from a sequence, then the probability of the left out data pair should be calculated using the estimated path, and the process should be repeated leaving out successive pairs of data values.

SUMMARY OF MO-MALCOM EXPERIMENTS

An encouraging outcome of both MALCOM and MO-MALCOM is that, after training, the estimated mean continuity map position for a given VQ code is highly correlated with the mean of the measured articulator positions that produce a the VQ code (Hogden, 1995; Nix, 1998). This is true even though the training data does not include articulator positions. This result was not wholly unexpected. In fact, statistical theory tells us that maximum likelihood parameters are typically (but see the discussion below) *consistent* (Neter, Wasserman & Kutner, 1985). That is, if our model is accurate then maximum likelihood parameter values will approach the actual parameter values of the system generating the data as the amount of training data gets large. Since the MO-MALCOM model parameters (the means and covariance matrices of the Gaussian probability density functions) constitute an estimate of the mapping between articulator positions and acoustics, the fact that the parameters are correlated with measured articulator positions suggests that the MO-MALCOM model has a great deal of validity.

Furthermore, Nix (1998) showed that MO-MALCOM positions are excellent at discriminating phonemes – better than measured articulator positions. Using a jackknife procedure, Nix used MO-MALCOM to create a CM from training data. Then, on testing data, smooth paths through the CM were found using only the VQ codes. Fisher's discriminant analysis was used to find dimension of the map that best discriminated the phoneme pair, where the phonemes in a pair are designated f_1 and f_2. Along this best dimension, the percentage of area in common between $p(\mathbf{x}|f_1)$ and $p(\mathbf{x}|f_2)$ was computed. To the extent that this is a low value, the CM positions give a lot of information about phoneme identity.

The ability of MO-MALCOM to differentiate between phonemes differing in place is demonstrated by two examples: 1) the largest overlap in MO-MALCOM PDFs between phoneme pairs composed of [p], [t], and [k] is 1%; 2) the largest overlap between phoneme pairs composed of [b], [d], and [g] is 6%. The ability of MO-MALCOM to discriminate between phonemes with similar articulation but different acoustics is also evident -- [b] and [p] have an overlap of less than 0.5%, [d] and [t] have an overlap of 2%, [k] and [g] have an overlap of 6%. Even [b] and [w] are discriminated well by MO-MALCOM positions (the overlap is less than 0.5%). Furthermore, MO-MALCOM continuity map positions are good at discriminating vowels -- the largest overlap for MO-MALCOM is 3% and only 6 vowel pairs have overlaps larger that 0.5%. The most difficult pair of phonemes for MO-MALCOM to discriminate are [r] and [l], which have 19% overlap. The next most difficult pair is [r] and the glottal stop with a 17% overlap. The vast majority of phoneme pairs have less than a 0.5% overlap and only 7 phoneme pairs have overlaps of more than 10%.

Despite good phoneme discrimination results, when MO-MALCOM was used to perform speaker dependent, isolated word recognition on data derived from the phonetically labelled portin of the switchboard data set, the recognition results were not impressive. Even on the training set, only about 40% recognition accuracy was achieved. However, there were many known deficiencies in the recognition system that was used (it was created in less than a year). First, the training set was much smaller than the speaker-independent continuous-speech recognition training sets commonly used today (we use about 3 minutes of speech as opposed to, say, 65 hours on the complete Switchboard training set). Second, doing isolated-word recognition, prevented the algorithm from taking advantage of a language model. Third, the model that estimates the probability of sequences of phonemes given a word was much more simplistic than in state-of-the-art recognition systems. Fourth, the dictionary contains only canonical pronunciations of words as opposed to pronunciations that commonly occur in casual speech. This problem is particularly severe since, in automatically extracting isolated words from continuous speech, phonemes were often added or deleted from the beginning or the end of the word. Sixth, we did not use cepstral mean subtraction or variance normalization.

DISCUSSION

The MO-MALCOM theory is still incomplete. While it is true that maximum likelihood parameters are typically consistent, particularly when using mixtures of Gaussians as with MO-MALCOM, it is not necessarily the case. Proofs that MO-MALCOM parameters actually are consistent would be welcome. If it can be proven that MO-MALCOM parameters are consistent, then it will be possible to argue that the mapping between acoustics and articulation (or possibly between acoustics and task-dynamic tract variable -- see below) can be recovered from acoustics alone. This would have important repercussions for the motor theory of speech perception as well as theories of speech production that postulate that phoneme targets must be acoustic because there is no teaching signal to help learn the mapping between acoustics and tract variables, e.g., (Guenther, Hampson & Johnson, 1998).

Furthermore, we are currently exploring variations on the MO-MALCOM theme, such as building a simplified task dynamic model (essentially constraining the paths to look like mass-spring motion between phone targets). Although not described above, simplifications to the MO-MALCOM model have been used to speed up processing. The effects of these simplifications on the results are unknown, and should be studied.

We believe that MO-MALCOM and its allies will prove to be valuable tools to add to our speech processing toolbox, and may well engender significant changes theories of speech perception and speech production.

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.

Bakis, R. (1991). Coarticulation modeling with continuous-state HMMs, Proceedings of the IEEE Workshop on Automatic Speech Recognition, (pp. 20-21). New York: Arden House.

Deng, L. (1998). A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. Speech Communication, 24, 299-323.

Guenther, F., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. Psychological Review, 105(4), 611-633.

Hogden, J. (1995,). A maximum likelihood approach to estimating speech articulator positions from speech acoustics. Paper presented at the Neural Information Processing Systems:Natural and Synthetic, Vail, Colorado.

Hogden, J. (1998). Phase 1 Final Report: An Articulatorily Constrained, Maximum Likelihood Approach to Speech Recognition (LA-UR 98-5638). Los Alamos, NM: Los Alamos National Laboratory.

Hogden, J., Lofquist, A., Gracco, V., Oshima, K., Rubin, P., & Saltzman, E. (1993). Inferring articulator positions from acoustics: an electromagnetic midsagittal articulometer experiment. Journal of the Acoustical Society of America, 94(3), 1764(A).

Jelinek, F. (1997). Statistical Methods for Speech Recognition. Cambridge, MA: MIT Press.

Lee, K. F. (1989). Automatic Speech Recognition: The Development of the SPHINX System. Boston: Kluwer Academic Publishers.

Martin, A., Fiscus, J., Przybocki, M., & Fisher, B. (1998, September 24-25). The Evaluation: Word Error Rates & Confidence Analysis. Paper presented at the Proceedings of the 9th Hub-5 Conversational Speech Recognition Workshop, Linthicum Heights, MD.

Neter, J., Wasserman, W., & Kutner, M. (1985). Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Design. (second edition ed.). Homewood, Illinois: Richard D. Irwin, Inc.

Nix, D. (1998). Machine learning methods for inferring vocal-tract articulation from speech acoustics. Unpublished Ph.D., University of Colorado, Boulder, CO.

Ostendorf, M., Digalakis, V., & Kimball, O. (1996). From HMM's to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition. IEEE Transactions on Speech and Audio Processing, 4(5), 360-377.

Picone, J., Pike, S., Regan, R., Kamm, T., Bridle, J., Deng, L., Ma, Z., Richards, H., & Schuster, M. (1999). Initial evaluatin of hidden dynamic models on conversational speech. International Conference on Acoustics, Speech, and Signal Processing, 1, 109-112.