LA-UR- 00-4807

Approved for public release; distribution is unlimited.

Title:	A Stochastic Articulatory-To-Acoustic Mapping as a Basis for Speech Recognition
Author(s):	John E. Hogden CCS-3 Patrick F. Valdez CCS-3
Submitted to:	IEEE conference on Measurement Technology Budpest, May 21-23, 2001

Los Alamos
NATIONAL LABORATORY

Form 836 (8/00)

DISCLAIMER

A STOCHASTIC ARTICULATORY-TO-ACOUSTIC MAPPING AS A BASIS FOR SPEECH RECOGNITION

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

Keywords: Speech Recognition, Speech Processing, Hidden-Dynamic Models.

Submitted for special session on multimodal recognition/synthesis

ABSTRACT

Hidden Markov models (HMMs) of speech acoustics are the current state-of-the-art in speech recognition, but these models bear little resemblance to the processes underlying speech production (Lee, 1989). In this respect, using an HMM to model speech acoustics is like using a Gaussian distribution to model data generated by a Poisson process – to the extent that the model is not an accurate representation of generating process, the accuracy of the model, and the meaning of the inferred parameters, is limited. Of This model mismatch likely contributes to the fact that state-of-the-art recognition performance (word accuracy) on recorded telephone conversations is only around 60-65%.

There have been recent attempts to create stochastic models of speech acoustics make more realistic assumptions about the mechanisms underlying speech production (Bakis, 1991; Deng, 1998; Hogden, 1998; Picone et al., 1999). In this paper we describe two stochastic models of speech production Conditional Observable Maximum Likelihood Mapping (CO-MALCOM) and its predecessor, Maximum Likelihood Continuity Mapping (MALCOM). The main component of both of these models is a stochastic mapping between speech acoustics and speech articulation.

A counter-intuitive aspect of the stochastic mapping is that the parameters of the mapping can be found using only acoustic data. While most speech researchers are familiar with the fact that HMM parameters can be estimated from acoustics alone, many still find it surprising that the mapping between speech acoustics and speech articulator positions (positions of the tongue, jaw, lips ...) can be found without articulator position measurements. Nonetheless, there are theoretical and experimental reasons to believe that MALCOM and it's allies learn a stochastic mapping between articulator positions and speech acoustics. Furthermore, CO-MALCOM can be combined with standard speech recognition algorithms to form a speech recognition approach based on a production model. Results of experiments related to MALCOM are summarized, and the CO-MALCOM extension is described.

BACKGROUND: STATE-OF-THE-ART SPEECH RECOGNITION

In many realistic domains, automatic speech recognition performance 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 unacceptable rate for many applications. Furthermore, recognition performance is not improving rapidly. 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 recognition results should prompt us to look for alternatives to the current approach.

The primary tools used in speech recognition are hidden Markov models (HMMs) -- 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, 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

Maximum Likelihood Continuity Mapping (MO-MALCOM) learns a stochastic model of sequences of categorical data values, e.g., vector quantization (VQ) codes representing speech acoustics. MALCOM assumes that 1) that data sequences are produced by objects moving smoothly through an abstract space called a continuity map (CM) and 2) the probability of observing a particular data value at time t is dependent on the position, x(t), of the object at time 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.

For pedagogical purposes, it is useful to think of the CM used by MALCOM as a mapping between acoustics and articulator positions. However, it is not necessary to believe that there is a one-to-one mapping between acoustics and articulator positions to believe that MALCOM can be useful. This is important because the extent to which articulator positions can be determine from a short-time window of speech acoustics is the subject of continuing debate (Atal, Chang, Mathews & Tukey, 1978; Hogden et al., 1993). Computer models and mathematical analyses have been used to argue that different vocal tract shapes can be used to produce the same acoustic signals. However, these theoretical analyses make many assumptions about speech production, as evidenced by the fact that the range of articulator positions that produce a given acoustic signal for one model may be an order of magnitude greater than the range of articulator positions that produce the same signal in a

[1] There was no 1999 evaluation and I have not yet been able to obtain results of the 2000 evaluation

different model. Furthermore, while the computer models may show that articulator positions can change by centimeters while still producing the same acoustic signal, relatively simple methods of recovering articulator positions from acoustics have errors of only around 1 or 2 millimeters. Regardless, for any acoustic speech signal, there must be some probability density function (PDF) quantifying the probability that articulator configuration x was used to create the speech signal. MALCOM approximates this fact using a parameterized distribution to indicate the probability of an articulator position at time t, $\mathbf{x}(t)$, conditioned on an observable representation of the acoustic signal, the VQ code at t, c(t), where $\mathbf{x}(t)$ is a vector and c(t) is a scalor. This density function is denoted $p[\mathbf{x}|c_i,\varphi]$, where φ is the set of parameters of the PDF (e.g. means and covariance matrices). Note that the density function is not necessarily unimodal, but that all the work described herein makes the simplifying assumption that the $p[\mathbf{x}|c_i,\varphi]$ is Gaussian.

While it may or may not be possible to invert a deterministic mapping from articulation to acoustics, 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,

$$P[c_i|\mathbf{x},\varphi] = \frac{p[\mathbf{x}|c_i,\varphi]P[c_i]}{p[\mathbf{x}|\varphi]} \qquad EQ.\ 1$$

Note that we are making an implicit conditional independence assumption here: the probability of outputting a VQ code is determined by the current articulator positions alone. The previous and subsequent articulator positions give no further information about the probability of the VQ code. This assumption is used to get the probability of a sequence of VQ codes, $\mathbf{c}$ = [c(1), c(2), ... c(n)] from a path through the CM, X:

$$P[\mathbf{c}|\mathbf{X},\varphi] = \prod_{t=0}^{n} P[c(t)|\mathbf{x}(t),\varphi], \qquad EQ\ 2$$

Clearly, if we had enough measurements of articulator positions and the resulting acoustics, a stochastic map like the continuity map could be made. However, articulator measurements are difficult to collect, so as a matter of practicality, the articulatory paths must be treated as unobservable. As discussed below, this does not preclude inferring parameters of the map, however. In the same way that HMM parameters can be inferred despite the fact that the state sequence is unobservable, the CM parameters can be inferred without being able to observe articulatory paths. However, to do so, we must put some constraints on the paths. The constraint that we have chosen is that only paths with no energy above some cut-off frequency are possible. This constraint is not only reasonable, since real articulator trajectories have nearly all their energy in low frequency components, but is particularly easy to implement.

To better understand the role of the smoothness constraint, consider the relationship between a smoothness constraint and a Markov constraint. Suppose that instead of the smoothness constraint, we had simply limited the time derivative of the articulatory trajectory. This would have been analogous to imposing a first-order Markov constraint in that we could look at any two sequential articulatory path positions to determine whether the trajectory was possible. So, with a constraint on the first derivative, given a starting articulator position, the probability of making a transition to

any position within some distance (determined by the derivative and the time between observations) would be some non-zero constant, but the probability of a transition to an articulator position outside that distance would be zero. In fact, the smoothness constraint does place limits on the derivative of the trajectory, since the derivative of the trajectory can not have any energy above the cut-off frequency. However, the smoothness constraint places limits on second derivatives, third derivatives, etc. Thus, imposing a smoothness constraint is more similar to using a high-order Markov model. Noting that the smoothness constraint only requires one parameter (the cut-off frequency), whereas the number of transition probabilities goes up exponentially for higher-order Markov models. So the smoothness constraint is not capable of representing the variety of phenomena that a Markov model can be used to model. Perhaps the smoothness constraint should be thought of as a high-order Markov constraint in which many of the transitions probabilities are tied. In any case, since real articulator trajectories are bandwidth limited, using a smoothness constraint instead of a Markov constraint provides a realistic and parsimonious way to limit articulatory trajectories.

As described below, the parameters are learned using maximum likelihood techniques. Statistical theory tells us that maximum likelihood estimates of mixture density parameters are *consistent* under relatively general conditions (McLachlan & Basford, 1988). That is, if the model reflects the underlying generating process, 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 constitute an estimate of the mapping between articulator positions and acoustics, we might expect that the MALCOM continuity map will approximate the actual stochastic mapping between acoustics and articulation given enough data. As discussed below, this appears to be the case.

MALCOM TRAINING

Signal Processing

Before applying 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 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.

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. Note that the two learning steps (given below) used to calculate the CM parameters are analogous to the 1) Viterbi algorithm as used to calculate the path through a HMM state space, 2) the HMM parameter re-estimation algorithm.

1) Given some initial set of PDF parameters, and many different examples of $\mathbf{c}_t$ VQ 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,

$$\hat{\mathbf{X}} = \arg\max_{\mathbf{X}} P[\mathbf{c}|\mathbf{X},\varphi] \qquad EQ.\ 3$$

2) Find the values of the PDF parameters that maximize the probability of the data sequence given the path estimates found in step 1. That is,

$$\hat{\varphi} = \arg\max_{\varphi} P[\mathbf{c}|\hat{\mathbf{X}},\varphi] \qquad EQ.\ 4$$

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

Optimizing Other MO-MALCOM Parameters

We also need to know the number of dimensions to use in the continuity map, and the cut-off frequency of the smooth paths. Trying many combinations of parameters to determine which combination works best for the problem being studied is the way to get the best performance. However, in many cases, determining the performance of the model on a task is much more time consuming than using MALCOM to estimate the probability of a cross-validation sample of sequences. In such a case, cross-validation is preferable. However, note that estimating 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 MALCOM path should be estimated without using one of VQ codes in a sequence, then the probability of the left out VQ code should be calculated using the estimated path, and the process should be repeated leaving out successive data values.

MALCOM EXPERIMENTS

Two studies showing that MALCOM accurately finds a mapping between acoustics and articulator positions will be reviewed (Hogden, 1995; Nix, 1998). Articulator data was not used for training MALOCM in either of these studies, although articulator measurements were used to calculate the extent to which MALCOM estimated articulator trajectories mimic actual articulator trajectories. Despite the fact that articulator positions were not used for training, correlations between estimated and actual articulator positions were in the 0.9 to 0.97 range for several important articulatory parameters.

MALCOM SPEECH RECOGNITION

For speech recognition, the task is to determine the probability of one sequence of categorical data values (e.g. phonemes) conditioned on an observable sequence of categorical data values. One modification of MALCOM, which we call Conditional Observable MALCOM (CO-MALCOM) is applicable to these types of tasks.

In CO-MALCOM, we start with the assumption that the articulator positions do not just tell us about the acoustics being output, but also tell us which phonemes are being produced. The process of recovering a phoneme sequence from a sequence of VQ codes is taken to be: 1) using a continuity map, find the articulator trajectory for a given sequence.; 2) given an articulator trajectory, find the probability of each phoneme at each time; 3) combine this information with word and language models to get the probability of a word given a VQ code sequence.

Estimating the probability of a phoneme sequence given an articulator trajectory is done using the same techniques used to estimate the probability of a VQ code sequence given an articulatory trajectory: 1) PDFs quantifying the probability of an articulator position conditioned on a phoneme are estimated; 2) Bayes' law is used to get the probability of a phoneme conditioned on an articulator position; 3) the probability for a whole sequence of phonemes conditioned on a trajectory is calculated using a conditional independence assumption.

Finding the path given the VQ code sequence is done slightly differently than described previously. Instead of finding the X that maximizes $P(\mathbf{c}|\mathbf{X},\varphi)$, we maximize $p(\mathbf{X}|\mathbf{c},\varphi)$ The justifications for this is pragmatic: there is an analytic solution to the problem of maximizing P(**X**|**c**) so we can quickly calculate the path, $\mathbf{X}(\mathbf{c},\varphi)$; 2) it is relatively easy to get the gradient of $\mathbf{X}(\mathbf{c},\varphi)$with respect to **c** and φ. These are important considerations, because they considerably reduce the complexity of finding the training process. The reduction in complexity is due to the fact that CO-MALCOM parameters should maximize $P[\mathbf{f}|\mathbf{X}(\mathbf{c},\varphi),\gamma]$, where **f** is the sequence of phonemes, and γ is the set of parameters giving the mapping between phonemes and articulation. Since this maximization requires using the chain rule to get the derivative of the probability with respect to φ, it is a tremendous savings to be able to calculate the derivative of **X** with respect to φ.

Combining CO-MALCOM with Word Models

In standard speech recognition algorithms, the probability of a phoneme sequence given a word, $P[f(t)=f_i|w]$, 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 this 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:

$$a_{ij} = P[f(t)=f_i|f(t-1)=f_j,w] \qquad EQ.\ 5$$

$$b_i(t) = P[f(t)=f_i|x(t)] \qquad EQ.\ 6$$

$$\pi_i = P[f(1)=f_i|w] \qquad EQ.\ 7$$

$$\alpha_i(t) = P[f(t)=f_j|w,x(t),x(t-1),...,x(1)] \qquad EQ.\ 8$$

Assuming conditional independence, the reader can confirm that

$$\alpha_i(1) = b_i(1)\pi_i \qquad EQ.\ 9$$

$$\alpha_i(t) = b_i(t)\sum_j a_{ij}\alpha_j(t-1) \qquad EQ.\ 10$$

Using these recursively calculated probabilities we can find

$$P[w|\mathbf{X}] = P[w|x(t),x(t-1),...,x(1)] = \sum_i \alpha_i(t)P(w) \qquad EQ.\ 11$$

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.

RELATED EXPERIMENTS

CO-MALCOM is an attempt to improve an earlier algorithm which we called MO-MALCOM, for Multiple Observable MALCOM. As such, it has not yet been evaluated on test data. Nonetheless, since CO-MALCOM eliminates flaws in MO-MALCOM, we expect performance to be on par with MO-MALCOM performance, or slightly better. For this reason, we will describe the results of a suggestive study MO-MALCOM. 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 the axis of the map that best discriminated the phoneme pair. Along this best dimension, the percentage of area in common between $p(\mathbf{x}|f_i)$ and $p(\mathbf{x}|f_j)$ 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 of articulation 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 labeled portion of the switchboard data set, the recognition results were not impressive (Hogden, 1998). 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), which leads us to believe that further tests are needed to assess recognition performance. First, the training set was much smaller than the speaker-independent continuous-speech recognition training sets commonly used today (we used 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 contained 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. Fifth, we did not use cepstral mean subtraction or variance normalization.

DISCUSSION

The CO-MALCOM and MALCOM theory is still incomplete. Since these techniques involve estimating mixture density parameters, it is reasonable to expect that CO-MALCOM parameters will be consistent. Nonetheless, a proof that CO-MALCOM parameters are consistent would be welcome. Although not described above, simplifications to the MALCOM model are used to speed up processing. The effects of these simplifications on the results are unknown, and should be studied. Furthermore, we are currently exploring CO-MALCOM variations, such as building a task dynamic model (Saltzman & Munhall, 1989) into CO-MALCOM.

We believe that MALCOM and its allies will prove to be valuable tools to add to our speech processing toolbox, and may well engender significant changes in 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. (1996). A maximum likelihood approach to estimating articulator positions from speech acoustics (LA-UR-96-3518). Los Alamos, NM: Los Alamos National Laboratory.

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.

Liberman, A., & Mattingly, I. (1985). The motor theory of speech perception revised. Cognition, 21, 1-36.

MacNeilage, P., Rootes, T., & Chase, R. (1967). Speech production and perception in a patient with severe impairment of somesthetic perception and motor control. Journal of Speech and Hearing Research, 10(3), 449-467.

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.

McLachlan, G. J., & Basford, K. E. (1988). Mixture Models: Inference and Applicatins to Clustering. (2 ed.). (Vol. 84). New York: Marcel Dekker, 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.

Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. Ecological Psychology, 1(4), 333-382.