LA-UR- 97 - 4848

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: **MODELING WORDS WITH SUBWORD UNITS IN AN
ARTICULATORILY CONSTRAINED SPEECH RECOGNITION
ALGORITHM**

AUTHOR(S): John Hogden

SUBMITTED TO: External Distribution - Hard Copy

RECEIVED
APR 06 1998
OSTI

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

MASTER

By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive royalty-free license to publish or reproduce the published form of this contribution or to allow others to do so, for U.S. Government purposes:

The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

Los Alamos

Los Alamos National Laboratory
Los Alamos New Mexico 87545

19980423 160

DTIC QUALITY INSPECTED 2

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, makes 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.

Modeling words with subword units in an
articulatorily constrained speech recognition algorithm

John Hogden
Los Alamos National Laboratory
11/20/97

RECEIVED
APR 06 1998
OSTI

The goal of speech recognition is to find the most probable word given the acoustic evidence, i.e. a string of VQ codes or acoustic features. Speech recognition algorithms typically take advantage of the fact that the probability of a word, w , given a sequence of VQ codes, c , can be calculated from:

$$p(w | c) = \frac{p(c | w)p(w)}{p(c)} \quad \text{EQ. 1}$$

Since the probability of a word, $p(w)$, is typically given by a language model, and the probability of a sequence of VQ codes, $p(c)$, is constant during recognition, our goal is to create a word model giving the probability of the sequence of VQ codes given the word.

Theoretically, MALCOM could find $p(c | w)$ using the relationship:

$$p(c | w) = \int p(c | X)p(X | w)dX \quad \text{EQ. 2}$$

Where X is a path through the continuity map (CM). If X is forced to be smooth we retain constraints that are analogous to requiring that the articulators move smoothly. However, since it is not practical to calculate the integral over all possible X values, a suboptimal technique for finding $p(c | w)$ must be employed. In such situations, one can use the maximum value of $p(c | X)p(X | w)$ as an estimate of $p(c | w)$. For example, given a sequence of acoustic codes and while requiring X to be smooth, we can find:

$$\hat{X} = \max_X [p(c | X)p(X | w)] \quad \text{EQ. 3}$$

and then use:

$$\hat{p}(c | w) = p(c | \hat{X})p(\hat{X} | w) \quad \text{EQ. 4}$$

as our estimate of $p(c | w)$.

One thing these equations make clear is that we should have some way of calculating $p(X | w)$ in order to perform speech recognition using an articulatorily constrained, maximum likelihood framework for speech recognition.

II. Subword modeling

There are reasons not to learn a model for each word: for large vocabulary speech recognition, it is difficult to get enough examples of every word in the vocabulary to adequately determine the word model parameters. So, to take maximal advantage of a set of training data, speech recognition algorithms commonly attempt to model subword units

(phonemes, diphones, etc.) instead of whole words. The subword units can then be concatenated to create word models. In contrast to creating a model for each of many thousands of words, only 40 or so phonemes are used in English. Thus, it is much easier to get adequate training data to determine the parameters of phoneme models than to determine the parameters of word models. With models for 40 or so phonemes and a dictionary that describes each word as an ordered set of phonemes, it is possible to create word models for words that have never even been observed. In this section, we describe an approach to using phoneme models of speech within an articulatorily constrained maximum likelihood framework. We will discuss the technique as if phoneme models are being used, although it might make more sense to use subphoneme units, e.g. treat /d/ as the sequence: /transition to d-closure/, /silence/, /d-burst/ ...

Recall that there is already some evidence that the positions in a CM made using MALCOM roughly correspond to articulator positions. This suggests that the CM positions corresponding to a given phoneme will be distributed in some relatively well-specified region of the CM. For example, suppose that one dimension (call it y) of the CM corresponds to lip rounding and another dimension (call it x) corresponds to the x position of the tongue tip. Since /t/ is always produced with the tongue tip forward but with varying degrees of lip rounding, we might expect the distribution of CM positions observed during the production of /t/ to have a small variance on the x -axis but a larger variance on the y -axis. Thus, a phoneme may be characterized by a distribution of positions in the CM.

Using $p(\mathbf{x} | w_i)$ to denote the distribution of positions in the continuity map given the phoneme w_i , and assuming conditional independence, we can write:

$$p(\mathbf{X} | \mathbf{w}) = \prod_i p(\mathbf{x}_i | w_i) \quad \text{EQ. 5}$$

The above equation does not take into account constraints on $\mathbf{X}$, which is the topic of the next section.

III. Aligning text to speech

Note that describing a word as an ordered sequence of phonemes is incomplete in that it leaves out information about the relative timing of the phonemes. This will cause a problem if we try to use EQ. 5 to aid in performing the constrained maximization of equation EQ. 3. Put simply, it is impossible to determine if $\mathbf{X}$ is a smooth path if we only know that $\mathbf{x}_1$ occurs before $\mathbf{x}_2$, $\mathbf{x}_2$ occurs before $\mathbf{x}_3$, etc. In order to fully characterize a particular production of a word we need to say not only what phonemes were produced, but when the phonemes were produced. However, it would not be a good idea to have a word model deterministically specify when the phonemes occur, because the timing of the phonemes will vary with speaking rate, emphasis, etc. Thus, in order to find the word that maximizes the probability of a sequence of VQ codes, we should use a probabilistic algorithm to find the alignment of the phonemes with the acoustics and to find the smooth path that, together with the alignment, jointly maximizes the probability of the VQ code sequence. The process outlined below will give 1) the probability of the acoustic sequence given a word, and 2) the alignment of the phoneme centers with the acoustics. As such, it is similar to using the viterbi algorithm to align text with speech.

Putting the problem into equation form, we need to first find:

$$[\hat{\mathbf{X}}, \hat{\mathbf{t}}] = \max_{[\mathbf{x}, \mathbf{t}]} [p(\mathbf{c} | \mathbf{X}, \mathbf{t}) p(\mathbf{X} | \mathbf{w}, \mathbf{t}) p(\mathbf{t})] = \max_{[\mathbf{x}, \mathbf{t}]} [p(\mathbf{c} | \mathbf{X}) p(\mathbf{X} | \mathbf{w}, \mathbf{t}) p(\mathbf{t})] \quad \text{EQ. 6}$$

where

$$\mathbf{t} = [\Delta t_1 \quad \Delta t_2 \quad \dots \quad \Delta t_M] \quad \text{EQ. 7}$$

and Δt_i is the time between the center of phoneme $i-1$ and phoneme i (the 0th phoneme is the beginning of the word). Then we can use:

$$\hat{p}(\mathbf{c} | \mathbf{w}) = p(\mathbf{c} | \hat{\mathbf{X}}) p(\hat{\mathbf{X}} | \mathbf{w}, \hat{\mathbf{t}}) p(\hat{\mathbf{t}}) \quad \text{EQ. 8}$$

to estimate $p(\mathbf{c} | \mathbf{w})$.

This approach to speech recognition will undoubtedly benefit from extended research into modeling $p(\hat{\mathbf{t}})$. After all, it is in $p(\hat{\mathbf{t}})$ that we could add information about how speaking rate transforms the relative timing of the phonemes. For example, we could make $p(\hat{\mathbf{t}})$ a function of speaking rate and constrain speaking rate to vary slowly over time. In addition, since $p(\hat{\mathbf{t}})$ varies by word, we either have to specify $p(\hat{\mathbf{t}})$ for each word (increasing the complexity of the word model) or we need to find a way to determine $p(\hat{\mathbf{t}})$ just from knowing the phoneme sequence. It would be possible to find $p(\hat{\mathbf{t}})$ from the phoneme sequence if, for example, the time between /a/ and /b/ is relatively independent of the word in which /a/ and /b/ are observed. However, our baseline model will be considerably simpler. Although we have not yet firmly decided on the baseline model, noting that the components of $\hat{\mathbf{t}}$ must all be greater than 0, we expect to model $p(\hat{\mathbf{t}})$ using Poisson distributions over the components of $\hat{\mathbf{t}}$.

Given $\hat{\mathbf{t}}$, it should be possible to find the $\hat{\mathbf{X}}$ in EQ. 3 using a maximization algorithm such as conjugate gradient ascent. While we have not yet derived the gradient equation, we feel that it will not be difficult. In contrast, we expect that it will be difficult to find the gradient with respect to the components of $\hat{\mathbf{t}}$. Therefore, it will be probably be necessary to specify a $\hat{\mathbf{t}}$, find $\hat{\mathbf{X}}$, record the value of $\hat{p}(\mathbf{c} | \mathbf{w})$, and repeat this process with different values of $\hat{\mathbf{t}}$. The $\hat{\mathbf{t}}$ for which $\hat{p}(\mathbf{c} | \mathbf{w})$ is maximized will be chosen. It should be noted that search techniques exist that can find maxima even without knowing gradients.

To review, $\hat{\mathbf{t}}$ gives the best alignment of the phoneme centers with the speech, and given the best alignment we use EQ. 4 to get an estimate of the likelihood of the acoustic sequence given the word. The likelihood of the acoustics sequence given the word can then be used with a language model to perform speech recognition.

IV. Summary

The steps we are proposing for training the speech recognition algorithm are:

- 1) Use unlabeled continuous speech data to make a CM for each speaker.
- 2) Using labeled isolated word data, find the CM positions corresponding to each phoneme, for each speaker.
- 3) For each phoneme, find a probability density function that characterises the CM map

positions corresponding to the phoneme.

- 3) For each word in the isolated word training data, find the distributions of Δt_i for each i . This distributions of $\hat{\mathbf{t}}$ combined with the phonetic transcription of each word will form the lexicon.

The steps we are proposing for testing the algorithm are:

- 1) Compare each word model in the lexicon to each word in the testing set. This will be done by:
 - a) For each word in the lexicon, find the $\hat{\mathbf{t}}$ and $\hat{\mathbf{X}}$ that jointly maximize $\hat{p}(\mathbf{c} | \mathbf{w})$.
 - b) Find the word in the lexicon with the maximum $\hat{p}(\mathbf{c} | \mathbf{w})$.
- 2) Calculate error rates.

M98004323

Report Number (14) LA-UR -- 97-4848

Publ. Date (11) 197711

Sponsor Code (18) DOE/DP, XF

UC Category (19) UC-705, DOE/ER

DOE