DESIGN OF A SPEECH RECOGNITION SYSTEM BASED ON ACOUSTICALLY DERIVED SEGMENTAL UNITS

M. Bacchiani[1] *M. Ostendorf*[2] *Y. Sagisaka*[1] *K. Paliwal*[3]

[1] ATR Interpreting Telecommunications Res. Labs. 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
[2] ATR ITL and ECS Engineering Deparment, Boston University, Boston, MA, USA
[3] ATR ITL and School of ME, Griffith University, Brisbane, Australia

ABSTRACT

The design of speech recognition system based on acoustically-derived, segmental units can be divided in three steps: unit design, lexicon building and pronunciation modeling. We formulate an iterative unit design procedure which consistently uses a maximum likelihood (ML) objective in successive application of resegmentation and model re-estimation. The lexicon building allows multi-word entries in the lexicon but restricts the number of these entries in order to avoid a too costly search. Selected multi-word lexical entries are those with high frequency (such as function words) and those which consistently exhibit cross-word phone assimilation. The stochastic pronunciation model represents the likelihood of a particular acoustic segment sequence given the phonetic baseform of a lexical item, where the sequence of baseform phones are treated as a Markov state sequence and each state can emit multiple segments.

1. INTRODUCTION

Great progress has been made in the development of recognition systems for continuous read speech, but the performance of these systems degrades severely when they are applied to spontaneous speech. This indicates that a different approach in modeling is required to design a system that is better suited to spontaneous speech. Our approach is to combine two advances proposed in previous work: the use of acoustically-derived subword units [1] and segmental modeling [2, 3]. In order to use such a modeling approach in a speech recognition system, we must solve three problems: 1) How do we design an inventory of acoustically-derived segmental units, 2) How do we construct a lexicon, and 3) How do we model the pronunciation of the lexical entries in terms of the acoustically-derived subword units? The use of segmental modeling allows us to exploit temporal correlation between successive frames of a given state [2]. Specifically, we use a vector polynomial function to model the segment mean trajectory. The resulting polynomial trajectory model is described in detail in section 2. Section 3 describes our approach to the unit design problem, the lexicon building problem is addressed in section 4, and section 5 describes how we model the pronunciation of the lexical entries in terms of the automatically derived units. Experimental results are provided in section 6, followed, in section 7, by a discussion of unresolved modeling questions.

2. POLYNOMIAL TRAJECTORY MODELS

In the polynomial trajectory model, we assume that successive cepstral features correspond to noisy observations of a smooth polynomial trajectory in vector space. More specifically, the observations $\underline{\mathbf{x}}_{i,j}$ in a segment $\underline{\mathbf{X}}_i = [\underline{\mathbf{x}}_{i,1}, \ldots, \underline{\mathbf{x}}_{i,N_i}]$ are conditionally independent and Gaussian given the segment length N_i and a length-dependent function for time-sampling the trajectory, e.g.

$$p(\underline{\mathbf{X}}_i|B_k, \Sigma_k) = \prod_{j=1}^{N_i} N(\underline{\mathbf{x}}_{i,j}; \mu_j(B_k), \Sigma) \quad (1)$$

where k is the model index and $N(\underline{\mathbf{x}}; \mu, \Sigma)$ denotes a vector Gaussian with mean μ, covariance Σ evaluated at $\underline{\mathbf{x}}$. Note that we assume a constant covariance for the duration of the segment. The mean trajectory is given by

$$[\mu_1 \cdots \mu_{N_i}] = \underline{\mathbf{B}}_k \cdot \underline{\mathbf{Z}}_{N_i} \quad (2)$$

where $\underline{\mathbf{B}}_k$ is a $D \times (R+1)$ matrix for an R-order polynomial, $\underline{\mathbf{Z}}_{N_i}$ is an $(R+1) \times N_i$ time sampling matrix, and

$$\underline{\mathbf{z}}_{N_i,j} = \left[1\, t_j\, t_j^2 \cdots t_j^R\right]^T; \quad t_j = (j-1)/(N_i - 1) \quad (3)$$

for $j = 1, \ldots, N_i$. The model parameters, $\underline{\mathbf{B}}_k$ and Σ, can be estimated using a maximum likelihood criterion as described in [3]. In addition to the Gaussian distribution on the observations, the segments are also characterized with a duration model $p(N_i|k)$.

3. UNIT DESIGN

The task of the unit design procedure is to define speech units consistent with some acoustic criterion and compute their models. Since we are using an R-th order polynomial trajectory model for segment modeling, the acoustic criterion we use for defining speech units assumes that the acoustic segments representing these units follow an R-th order polynomial. Our unit design procedure is iterative in nature and can be outlined as follows:

1. *acoustic segmentation*
2. *initial clustering*
3. *iterative re-estimation*
   (a) *maximum likelihood segmentation*
   (b) *maximum likelihood parameter estimation*
   (c) *go to 3a*

Since the models resulting from the unit design procedure are used in a likelihood based recognition system, we consistently use maximum likelihood as the design criterion in all steps of the algorithm. The iterative unit design is similar to the one for segment quantizer design in [4], the major difference being that our objective is maximum likelihood rather than minimum distortion. The different steps in the algorithm are described in detail below.

0-7803-3192-3/96 $5.00©1996 IEEE

3.1. Acoustic segmentation

The acoustic segmentation step functions as an initialization of the unit design procedure. Taking an approach similar to that in [1], the maximum likelihood (ML) segmentation of the training data is found by use of dynamic programming (DP). Segmentation is done under the assumption that "segments" of speech (not necessarily phones) are coherent in the sense of following an R-th order polynomial trajectory model. For example, speech is piecewise constant for $R = 0$ and piecewise linear for $R = 1$. The likelihood of the segments during the DP is computed using a multivariate Gaussian model with a single diagonal covariance used for all the segments. As no model inventory is available at this point, model parameters are fit to the individual hypothesized segments considered in the DP search. The very limited amount of data found in such a segment often prohibits the estimation of an invertible covariance matrix. To circumvent this problem, a single diagonal covariance matrix, estimated from a full utterance, is used.

During segmentation, the likelihood increases monotonically with the increase of the number of allowable segments. We control the average number of segments by setting a fixed threshold on the average likelihood per frame.

The difference between the work described in [1] and the work described here is that we use a multivariate Gaussian model to compute segment likelihoods instead of using a Euclidean distance measure.

3.2. Initial clustering

The segments resulting from the acoustic segmentation are clustered to form an initial inventory of acoustic segmental units. The clustering objective is maximum likelihood, which can be equivalently implemented as a "multivariate Gaussian distance measure":

$$N \log(|\Sigma|) + \mathrm{tr}\left\{N S \Sigma^{-1}\right\} + (\underline{\mathbf{B}}_s \underline{Z}_N - \underline{\mathbf{B}}_c \underline{Z}_N)^T \Sigma^{-1} (\underline{\mathbf{B}}_s \underline{Z}_N - \underline{\mathbf{B}}_c \underline{Z}_N), \quad (4)$$

where $\mathrm{tr}\{\cdot\}$ denotes the trace operation, T denotes a matrix transpose, S is the sample covariance of a segment which has a trajectory regression matrix $\underline{\mathbf{B}}_s$, and every cluster is represented by a cluster mean trajectory $\underline{\mathbf{B}}_c$ and frame-level covariance Σ.

The clustering is performed using a likelihood-based K-means clustering algorithm, initialized by the models resulting from a binary tree (divisive) clustering algorithm. The desired number of clusters C is fixed heuristically beforehand. Let K^p be the number of clusters at iteration p. The binary tree clustering algorithm can be outlined as follows:

1. Initialize: Set $p = 0$, $K^0 = 1$ and compute the parameters corresponding to assigning all data to one cluster.
2. Select the best cluster to split, Y^p with parameters y^p, such that
   - the average likelihood per frame P_{ave} is smaller than all other clusters, and
   - the number of frames in the segments contained in the cluster exceeds a fixed threshold F.
3. Split this cluster into Q_1^p and Q_2^p, iterating to find new cluster representative (or model parameters):
   (a) Initialize the two cluster parameters with $q_1^p = y^p$ and q_2^p as a perturbed version of y^p.
   (b) Reassign the segments in Q_1^p and Q_2^p to the closest cluster according to the multivariate Gaussian distance measure, and compute the new average likelihood per frame P_{new}.
   (c) Go to 4 if the change in likelihood is small
   $$(P_{new} - P_{ave})/P_{ave} < \epsilon \quad (5)$$
   (d) Recompute the cluster representatives q_1^p and q_2^p using the ML criterion as described in [3].
   (e) Set $P_{ave} = P_{new}$ and go to 3b.
4. Increment the number of models K^p. If the desired codebook size is reached, $K^p = C$, then stop. Otherwise, increment p and go to 2.

After the algorithm above terminates, C cluster representatives are found and these are subsequently used to initialize the K-means clustering algorithm. The C cluster representatives that are found after termination of the K-means clustering algorithm form the models for the inventory of acoustically-derived segmental units. The subsequent steps in the unit design algorithm can be used to further refine the inventory of models.

If during the K-means clustering algorithm a cluster is found with less than F frames contained in the segments in the cluster, the cluster is removed from the inventory. The segments previously contained in this cluster are assigned to the cluster whose cluster representative is nearest in terms of the multivariate Gaussian distance measure.

3.3. Iterative re-estimation

Since the W segments resulting from the acoustic segmentation are now described by an inventory of C units with $W >> C$, the segment boundaries for the model-inventory will be sub-optimal. Let $\mathcal{M}$ denote the inventory of models associated with the C units. One can obtain the optimal segmentation, given the inventory $\mathcal{M}$, by performing a Viterbi segmentation using the Gaussian segment models. The Viterbi segmentation simultaneously finds optimal segment boundary positions and segment identities given the inventory of models. Let N_{max} be the maximum allowed length of a segment. The DP search computes for each frame in a T frame utterance:

For $t = 1, \ldots, N_{max}$:

$$\delta_1(i) = \log[p(y_1, \ldots, y_t | l, i) p(l|i) p(i)]; \ \forall i \in \mathcal{M}, l = t \quad (6)$$

For $t = 2, \ldots, T$:

$$\delta_t(i) = \max_{j \in \mathcal{M}, \tau \leq t} (\delta_{\tau-1}(j) + \log[p(y_\tau, \ldots, y_t | l_\tau, i) \cdot p(l_\tau|i) p(i|j)]) \ ; \ \forall i \in \mathcal{M}, \ l_\tau = t - \tau + 1 \quad (7)$$

$$\psi_t(i) = \operatorname*{argmax}_{j \in \mathcal{M}, \tau \leq t} (\delta_{\tau-1}(j) + \log[p(y_\tau, \cdots, y_t | l_\tau, i) \cdot p(l_\tau|i) p(i|j)]) \ ; \ \forall i \in \mathcal{M}, \ l_\tau = t - \tau + 1 \quad (8)$$

where y_t denotes the t-th frame in the utterance. To limit the cost of such a DP search, the values of τ are further constrained so that the number of segments for time t can only change with a fraction of the number of segments for time t in a previous iteration.

The observation probability $p(y_\tau, \cdots, y_t | l_\tau, i)$ denotes the likelihood that frames y_τ to y_t were emitted by model i given that the length of the segment is l_τ frames. The length

probability term $p(l_\tau|i)$ denotes the likelihood that model i produces an l_τ length segment. The unigram probability $p(i)$ and bigram probability $p(i|j)$ denote the probability of model i and i in the context of model j respectively. The optimal segmentation can now be derived by determining:

$$m_T = \underset{i \in \mathcal{M}}{\operatorname{argmax}}\, \delta_T(i) \tag{9}$$

and then trace back to the beginning of the utterance using the $\psi_t(i)$ values. The log-likelihood of the observations is computed as:

$$p(y_1, \cdots, y_N|l_N, i) = -\frac{1}{2}ND\log(2\pi) - \frac{1}{2}N\log(|\Sigma_i|) - \frac{1}{2}\sum_{m=1}^{N}(y_m - \underline{\mathbf{B}}_i\underline{\mathbf{z}}_{N,m})^T\Sigma_i^{-1}(y_m - \underline{\mathbf{B}}_i\underline{\mathbf{z}}_{N,m}) \tag{10}$$

where Σ_i denotes the covariance matrix of the i-th model, $\underline{\mathbf{B}}_i$ denotes the mean trajectory parameters of the i-th model and $\underline{\mathbf{z}}_{N,m}$ denotes the time-dependent polynomial terms defined in equation 3.

The length probability distribution is estimated by smoothing the distribution, computed from the relative frequencies of observed segment lengths. The model unigram and bigram distributions are estimated likewise.

Unless the model inventory and segmentation are both optimal in terms of the ML objective, the resegmentation will change the position of segment boundaries. Therefore, the likelihood can be improved by ML re-estimation of the model inventory. As all segments have a label, model re-estimation can be performed by computing the ML parameter estimate of the segments with corresponding label using the approach outlined in [3].

4. LEXICON BUILDING

The unit design algorithm will not necessarily produce segment boundaries at word boundaries. This problem can be prevented by imposing boundary constraints at word boundaries for the unit design process but this might limit the potential of the resulting acoustically-derived segmental units. Another solution is to allow multi-word entries in the lexicon, but when all possible multi-word entries are included in the lexicon, the lexicon size becomes so large that the search will become too costly. Therefore, the following algorithm is used to construct a lexicon which includes some multi-word lexical entries but limits the total number of entries to restrict the search cost.

1. Build an initial lexicon, consisting of all single-word entries and multi-word entries found in the training data with segments spanning all word boundaries within the sequence. A segment spans a word boundary if the time difference from the word boundary to the closest segment boundary is more than T ms.
2. For each multi-word boundary, compute a score from the time difference between the word boundary and the closest automatic boundary.
3. Prune the multi-word lexical entries with a low score until the desired lexicon size L is reached by adding boundary constraints at word boundaries.
4. Resegment using the model inventory, retaining all derived boundary constraints and re-estimate the models as described in section 2.

The score in step 2 is used as a measure of cross-word phonetic assimilation, and is computed as:

$$S(e) = \frac{1}{I(e)}\sum_{i \in \{e\}}\sum_{c \in B(i)}|c - A(c)| \tag{11}$$

where $S(e)$ denotes the score for lexical entry e, $I(e)$ denotes the number of word boundaries within lexical entry e (i.e. a 3-word sequence would have $I(e) = 2$) , $\{e\}$ denotes the instances of lexical entry e in the training data, $B(i)$ denotes the phone-based word-boundary times of instance i and $A(c)$ denotes the nearest edge of an automatically derived segment to time c. Note that the score includes a bias toward frequent multi-word sequences.

By applying this algorithm, we obtain a lexicon consisting of all distinct single word entries found in the training data plus the most frequently reduced word sequences which are better modeled by a multi-word lexical entry. Optionally, we can reduce the size of the lexicon in stages, repeating steps (2)-(4) to minimize the changes in the acoustic units at each stage.

5. PRONUNCIATION MODELING

In our pronunciation modeling we construct a stochastic model which estimates the likelihood of observing a sequence of automatically derived units given a phonemic baseform. The idea to derive a pronunciation from a phonemic baseform is very similar to the work described in [5, 6] although we attempt to model pronunciation in terms of automatic units rather than phones. Let ϕ_1^N denote the baseform pronunciation of a word, which is given by a phonetic dictionary, and U_1^M an automatic unit label sequence. The probability of a particular automatic label "pronunciation" U_1^M and baseform pronunciation ϕ_1^N is then

$$\begin{aligned} P(U_1^M, \phi_1^N) &= \sum_{\{\alpha_1^N\}} P(U_1^M, \phi_1^N, \alpha_1^N) \\ &= \sum_{\{\alpha_1^N\}} P(\alpha_1^N|\phi_1^N)P(\phi_1^N) \end{aligned} \tag{12}$$

where α_i is the (possibly empty) subsequence of U_j's that are associated with ϕ_i, and α_1^N contains all the information in U_1^M. The set of all possible partitionings of U_1^M into subsequences is denoted by $\{\alpha_1^N\}$. Assuming that the baseform sequence ϕ_1^N is Markov and that the α_i's are conditionally independent given the "states" ϕ_i, the probability of a particular observed pronunciation of word ϕ_1^N is

$$P(U_1^M, \phi_1^N) = \sum_{\{\alpha_1^N\}}\prod_{i=1}^{N} P(\alpha_i|\phi_i)P(\phi_i|\phi_{i-1}). \tag{13}$$

In [6], $P(\alpha|\phi)$ is represented separately for all possible subsequences α, since the total number is small (the observed subsequence lengths $l(\alpha)$ are limited to be ≤ 2 with additional restrictions on the number of length 2 sequences). In our case, on the other hand, $l(\alpha)$ can vary substantially, so some simplifying assumptions are needed to reduce the parameter space. One solution, implemented in our initial work, is to use phone-dependent unit label n-grams, e.g.

$$P(\alpha|\phi) = P(l(\alpha)|\phi)P(U_{t(\alpha)}|\phi)\prod_{j=t(\alpha)+1}^{t(\alpha)+l(\alpha)-1} P(U_t|U_{t-1}, \phi) \tag{14}$$

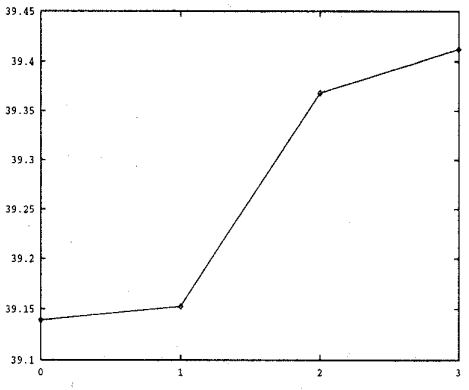

Figure 1. Likelihood as function of iterations on testing data.

where $t(\alpha)$ is the time of the first acoustic unit label in α and we use the fact that $P(\alpha|\phi) = P(\alpha, l(\alpha)|\phi)$.

6. EXPERIMENTAL RESULTS

To investigate the properties of the acoustic unit design algorithm, we conducted a number of experiments on the TIMIT database, using the male speakers from the New-England dialect-region (24 speakers for training, 7 for testing, 10 sentences per speaker). We generated 16 dimensional LPC derived cepstral feature vectors for the data using a 25.6 ms Hamming window with a 10 ms frame rate. An energy coefficient was appended to the vectors. The increase of the average likelihood per frame on the test set, using an inventory size of 300 zeroth order full covariance models is depicted in figure 1, which suggests that only a few iterations of retraining are needed. The likelihood of iteration 0 is the likelihood of the models obtained after the initial clustering step.

To compare the performance of the automatically-derived acoustic units versus phonetic units we constructed a set of allophonic unit models starting from the TIMIT phone segmentations. We separately clustered the segments corresponding to each phonetic label in the same way as for the acoustically-derived units described in section 2. The stopping criterion used for each cluster was that the average likelihood per frame was higher than some threshold P or that the number of frames was lower than some threshold M. We obtained 277 zeroth order allophonic unit models with diagonal covariance for 48 phone labels in this way. As a comparison, we computed an inventory of 277 acoustically-derived unit models with diagonal covariance and compared the likelihood, performing a Viterbi segmentation of the test set. The test set likelihood using the acoustically-derived unit models was higher than that based on the allophonic unit models given by the hand-labeled phones, which suggests that the acoustically-derived units fit the data better.

It has already been shown that higher order segment models outperform the zeroth order model in TIMIT vowel classification experiments [3], but we confirmed these results in our own vowel and phone classification experiments.

We tested the lexicon building algorithm by constructing a lexicon based on the automatic segmentations obtained by using 292 zeroth order models with diagonal covariance for all of the TIMIT corpus. The data included 326 male and 136 female speakers and 8 utterances per speaker, giving a total number of 4891 distinct single word entries in a corpus of 30132 non-unique words. Using a 20 ms threshold for step 1 of the algorithm described in section 3, 3422 distinct multi-word sequences were found. Multi-word sequences were pruned until a lexicon of size 5000 was obtained. The multi-word sequences were typically function word combinations, such as "in the", "do you", "of the", "on the", etc., as one might expect. Almost all additional lexical items were two word sequences. Many of the multi-word sequences that were not included were singleton examples of that word sequence.

7. DISCUSSION

In this paper, we have proposed a method for designing acoustically-derived segmental units, a lexicon containing multi-word entries for representing cross-word phonetic reduction/assimilation, and the mapping between a phonetic baseform associated with the lexicon and the observed segmental units. Preliminary experiments on the TIMIT corpus demonstrate the feasibility of each step, but it remains to be shown that the different components together offer a significant advantage over phonetic lexical modeling in continuous word recognition. However, we believe that using units that can be longer than phones and potentially span word boundaries will be important for handling the reduction phenomena that occurs in spontaneous speech.

Two main issues are still to be addressed for obtaining good word recognition results. First, the model is unlikely to provide significant gains for piecewise constant segment means, but efficient search algorithms are needed for practical implementation of polynomial trajectories of higher order. Second, the simplifying assumptions made in the pronunciation model, equations 13 and 14, are too restrictive, and we plan to apply decision tree clustering techniques similar to those of [6] to enable conditioning on a broader context of phone units.

REFERENCES

[1] K.K. Paliwal, "Lexicon building methods for an acoustic sub-word based speech recognizer," *Proc. of the Int. Conf. on Acoust., Speech and Signal Proc.*, pp. 729-732, 1990.

[2] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Trans. on Acoust., Speech and Signal Proc.*, vol. 37, no. 12, pp. 1857-1869, 1989.

[3] H. Gish and K. Ng, "A segmental speech model with applications to word spotting," *Proc. of the Int. Conf. on Acoust., Speech and Signal Proc.*, vol. II, pp. 447-450, 1993.

[4] Y. Shiraki and M.Honda, "LPC speech coding based on variable-length segment quantization," *IEEE Trans. on Acoust., Speech and Signal Proc.*, vol. 36, no. 9, pp. 1437-1444, 1988.

[5] J. M. Lucassen and R. L. Mercer, "An information theoretic approach to the automatic determination of phonemic baseforms," *Proc. of the Int. Conf. on Acoust., Speech and Signal Proc.*, vol. III, pp. 42.5.1-42.5.4, 1984.

[6] M. Riley, "A statistical model for generating pronunciation networks," in *Proc. of the Int. Conf. on Acoust., Speech and Signal Proc.*, vol. II, pp. S11.1-S11.4, 1991.