Unsupervised Learning of Non-Uniform Segmental Units for Acoustic Modeling in Speech Recognition

M. Bacchiani, M. Ostendorf[1], Y. Sagisaka and K. Paliwal[2]
ATR Interpreting Telecommunications Res. Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

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 (non-uniform) subword units and segmental modeling. Two questions must be answered when this approach is chosen: 1. How does one derive an inventory of acoustic units? and 2. How does one map these acoustic units to a lexicon? In the work presented here, we attempt to formulate an answer to the first question.

2. Iterative acoustic segmental unit design

As an initial step, we derive an inventory of acoustic segmental units by clustering speech segments that are obtained by acoustic segmentation, taking an approach similar to that in [1] to find the maximum likelihood segmentation of the training data by use of dynamic programming (DP). 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. A difference between the work described in [1] and the work described here is that we not only use zeroth order multivariate Gaussians (i.e. HMM's) but also higher order regression models as described in [2]. The resulting segments are clustered to obtain an initial inventory of acoustic segmental units. As we aim to use the resulting inventory of units in a Gaussian based recognition system, we use a multivariate Gaussian "distance measure" in clustering the segments, i.e. a maximum likelihood (ML) clustering criterion which represents each cluster with a mean trajectory and a frame-level covariance matrix estimated as in [2]. The (heuristically determined) number of clusters C is obtained by binary splitting the cluster with the lowest likelihood per frame given that the cluster holds more than a minimum number of frames. When the desired number of clusters is obtained by this method, the K-means clustering algorithm is applied, again with an ML criterion, using the C cluster representatives. If during this process a cluster with less frames than the set threshold is found, the cluster is removed from the inventory and the segments contained in it are re-assigned.

Since the N segments resulting from the acoustic segmentation are now described by an inventory of C units with $N >> C$, the segment boundaries for the model-inventory will be sub-optimal. One can obtain the optimal segmentation, given the inventory of units, by performing a Viterbi segmentation using the Gaussian segment models. After obtaining this segmentation, the models can be re-estimated using the maximum likelihood objective. The models can be improved further by iterating the re-segmentation and re-estimation steps. This iterative retraining algorithm is similar to the one described for segment quantizer design [3], the major difference being that our objective is maximum likelihood rather than minimum distortion.

3. Experiments

To investigate the properties of our learning algorithm, we conducted a number of experiments on the TIMIT database, using the male speakers from dialect-region 1 (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 during the iterative unit design, using an inventory size of 300 zeroth order full covariance models is depicted in figures 1 and 2, which suggest that only a few iterations of

[1]Current address: Boston University, Department of ECS Engineering, 44 Cummington Street, Boston, MA 02215
[2]Current Address: School of Microelectronic Engineering, Griffith University, Brisbane, QLD 4111, Australia

retraining are needed. The likelihood of iteration 0 is the likelihood of the models obtained after clustering the segments derived by acoustic segmentation. The half iterations (0.5, 1.5 and 2.5) in figure 1 correspond to the likelihood after re-segmentation, the full iterations (1, 2 and 3) correspond to the likelihood after re-estimation.

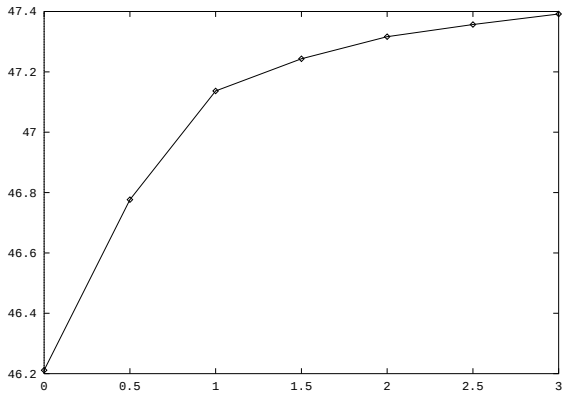

Figure 1. Likelihood as function of iterations on training data.

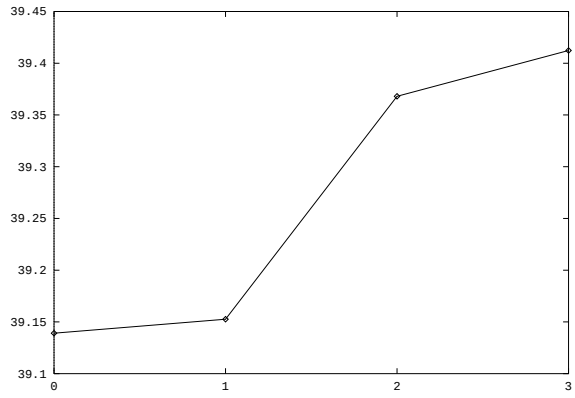

Figure 2. Likelihood as function of iterations on testing data.

To compare the performance of the automatically-derived acoustic units versus phonetic units we constructed a set of allophonic models starting from the TIMIT phone segmentations. We separately clustered the segments corresponding to each phonetic label in the same way as for the non-uniform 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 models with diagonal covariance for 48 phone labels in this way. As a comparison, we computed an inventory of 277 acoustic unit models with diagonal covariance and compared the likelihood, performing a Viterbi segmentation of the test set. The likelihood using the non-uniform unit models was 791199 compared to 775284 using the allophonic models. It has already been shown that higher order segment models outperform the zeroth order model in TIMIT classification experiments [2], but we confirmed these results in our own vowel and phone classification experiments.

4. Conclusion

In summary, this work has proposed the use of acoustically-derived segmental subword units and described an iterative design process for learning the units and associated parameters. Initial experiments show that the acoustically derived units offer the potential for improved performance over phonetic units in that they yield an increase in test set likelihood for equivalent numbers of free parameters. In addition, phone classification experiments also confirmed that higher order models are better able to capture spectral dynamics. In order to show similar (and hopefully greater) gains in recognition with the combination of higher order models and non-uniform units, the next step in the research is to address the problem of lexical mapping.

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] 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.

[3] 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.