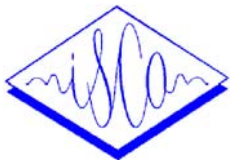

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

SEGMENT BOUNDARY ESTIMATION USING RECURRENT NEURAL NETWORKS

Toshiaki Fukada Sophie Aveline Mike Schuster Yoshinori Sagisaka

ATR Interpreting Telecommunications Research Laboratories
2–2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Tel: +81 774 95 1301, FAX: +81 774 95 1308, E-mail: fukada@itl.atr.co.jp

ABSTRACT

This paper describes a segment (e.g. phoneme) boundary estimation method based on recurrent neural networks (RNNs). The proposed method only requires acoustic observations to accurately estimate segment boundaries. Experimental results show that the proposed method can estimate segment boundaries significantly better than an HMM based method. Furthermore, we incorporate the RNN based segment boundary estimator into the HMM based and segment based recognition systems. As a result, the segment boundary estimates give useful information for reducing computational complexity and improving recognition performance.

1. INTRODUCTION

Accurately estimating segment boundaries is one of the most important techniques in (1) automatic segmentation [1][2] for acoustic model training and in (2) preprocessing for segment based speech recognition [3]. Conventional segmentation algorithms attempt to locate optimal segment boundaries either by minimizing distortion metrics through dynamic programming based methods [1] or by maximizing the metric score of acoustic models [2]. These algorithms, however, require acoustic (and language or duration) models to obtain adequate results. Nevertheless, even with such models, the estimated results are generally still poor because the approaches are not designed to detect boundaries, but rather to minimize or maximize scores for acoustic observations (e.g. cepstrum). Neural networks (NNs) that show a high performance for many classification tasks are suitable for estimating accurate boundaries. There have recently been several reports that boundary information obtained from feed-forward multilayer perceptrons (MLP) improves recognition performance [4][5].

In this paper, we propose a segment (e.g. phoneme) boundary estimation method based on bi-directional recurrent neural networks (BRNNs). A BRNN can be trained without the limitation of using a fixed size input window, and it gave better classification performance than a regular RNN on test problems [6]. The proposed method only requires acoustic observations to estimate segment boundaries, and networks are trained to accurately detect segment boundaries. We apply segment boundary estimation

1. to improve recognition performance using the network outputs and
2. to reduce computational complexity of segment based recognition using estimated candidates.

2. BRNN BASED SEGMENT BOUNDARY ESTIMATION

2.1. BRNN Structure

Bi-directional Recurrent Neural Networks (BRNNs) [6] are used for segment boundary estimation, and their structure is illustrated in Fig. 1. BRNNs can recursively accommodate forward and backward inputs to predict current output by only using a single network. A conventional RNN only uses input information from one side for the currently estimated output.

2.2. Input and Output

Feature parameter vectors (e.g. cepstrum) are used for the BRNN input, and the outputs (target values) are chosen according to whether the current frame is a boundary (out=1) or not (out=0). Figure 2 shows an example of outputs for BRNN based segment boundary estimation. The dotted line represents a true (target) output and the solid line represents an estimated output. These results were obtained for open test data using the network trained as described in **3.1.**.

2.3. Segment Boundary Estimation Algorithm

To determine segment boundaries from the BRNN outputs as shown in Fig. 2, the following three methods are used. A certain time point (frame) t is said to be a boundary if:

1. the output at t is above threshold h and is a local maximum;
2. the output at t is above threshold h or is a local maximum between a lower threshold $l(< h)$ and h;
3. the same as method 2, but for segment boundaries whose outputs are above threshold h, only every k-th time point is taken.

Method 1 is the simplest method and can be directly used to determine segment boundaries. Methods 2 and 3 are a little bit more complicated and possibly involve

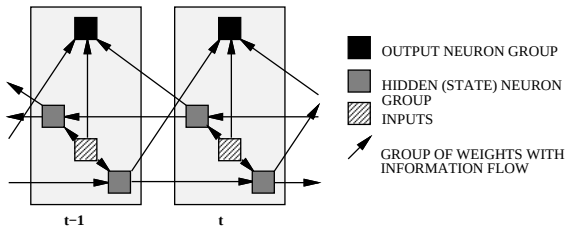

Figure 1. Bi-directional recurrent neural networks.

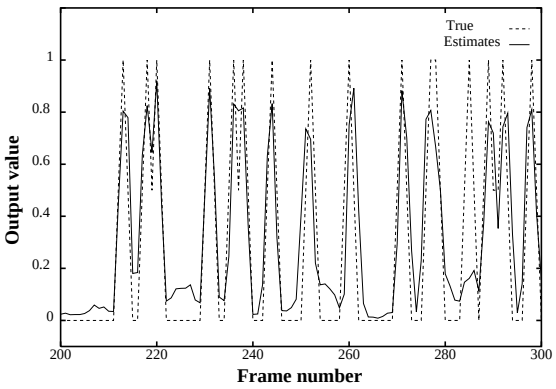

Figure 2. An example of BRNN outputs.

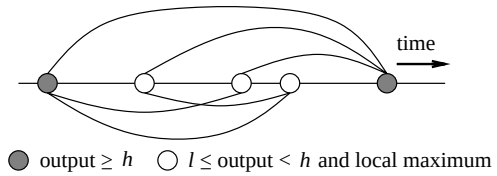

Figure 3. Lattice representation of segment boundary candidates.

first constructing a segment lattice (Fig.3) to characterize possible boundaries. Next, this segment lattice can be pruned by using more sophisticated evaluation methods (e.g. segment model based phoneme recognition) than simple network outputs to extract "better" boundaries.

3. PHONEME BOUNDARY ESTIMATION EXPERIMENTS

To investigate whether the proposed methods are useful, (1) a comparison between estimated boundaries obtained by method 1 and HMM based phoneme recognition results, (2) a comparison between BRNN and MLP and (3) a comparison among methods 1, 2 and 3 were done using the TIMIT database.

3.1. Conditions

26-dimensional MFCCs (12-dimensional MFCC + power and their derivatives) computed with a 25.6 msec window duration and a 10 msec frame period were used as the BRNN inputs. Based on the phoneme label information of the database, 1.0 is given for outputs if the current frame is a segment (phoneme) boundary, 0.5 is given if the current frame is right next to the boundary, and for anything else 0 is given. 10 forward and backward states and 30 hidden states for the BRNN were used (2,181 weights in total), and BRNN training was done using 462 speakers with 1,000 iterations. Test data was used for 168 speakers (50,318 boundaries, about 410,000 frames). The mean squared errors between true (target) and estimates were 0.0604 for training data and 0.0621 for testing data, respectively. Thresholds in methods 1, 2 and 3 were experimentally set to $h = 0.4$ and $l = 0.1$.

To compare the proposed method with the HMM based method, context-independent (CI) models, context-dependent (CD) models and a phoneme bigram language model were generated for 61 TIMIT phonemes. Left-to-right HMMs with 3 states for each phoneme and 5 Gaussian mixture components per state were trained for the CI models. As for the CD models, shared-state HMMs (600 states in total) with 3 Gaussian mixture components per state were trained [7]. The feature parameters and the training data were the same as for the BRNN conditions.

The MLP structure was tested here for three different structures allowing the use of the following three amounts of acoustic context: (1) one frame as input (MLP-1), (2) three frames (middle, left and right) as input (MLP-3), and (3) five frames (middle, two left and two right) as input (MLP-5). The structures of these networks were adjusted so each of them had about the same number of free parameters for the BRNN (approximately 2,200 here). The feature parameters, the training data, the iterations and the thresholds were the same as for the BRNN conditions.

3.2. Evaluation Criterion

To evaluate the estimated results, $Correct = H/N \times 100(\%)$ and $Accuracy = (N - D - I)/N \times 100(\%)$ were used, where

- H (Hit) : estimated boundary was within a $\pm M$ frame margin of the true boundary
- D (Deletion) : no estimated boundary was within $\pm M$
- I (Insertion) : estimated boundary was not within $\pm M$
- N : total number of true boundaries ($N = H + D$).

Note that if several estimated boundaries i were within $\pm M$, $i - 1$ were treated as insertions.

3.3. Results

3.3.1. Comparison between method 1 and an HMM based approach

Estimation results with margins of $M = 0, 1, 2$ are shown in Table 1. To produce reference results, an evaluation was performed by using the boundaries obtained through HMM based phoneme recognition with HMMs and a phoneme bigram language model. The results for the context-independent HMMs and the context-dependent HMMs are listed in Table 2(a) and Table 2(b), respectively. Comparing Table 1 with Table 2, the proposed method gives considerably higher accuracy than the HMM based approach, especially for $M = 0$ or $M = 1$, even though the BRNN based approach does not use any linguistic knowledge. The reason might be that the BRNNs are trained to accurately detect segment boundaries, while the HMMs are trained based on maximum likelihood criteria.

3.3.2. Comparison between BRNN and MLP

Table 3 shows the comparison of estimation performances between BRNN and MLP for method 1. The BRNN structure results in the best performance. Moreover, it has the advantage that one does not have to choose the optimum number of consecutive frames to define an input window size.

3.3.3. Comparison among methods 1, 2 and 3

Estimation results for methods 1, 2 and 3 are shown in Table 4. Skip step k in method 3 and M were both set to 2. Method 1 gave the highest accuracy, but there were a large number of deletion errors. This indicates that method 1 would not be appropriate when boundary candidates could be evaluated with other techniques as

Table 1. Estimation results based on the BRNN (method 1).

	Margin		
	0	1	2
Hit	23,175	38,248	40,056
Deletion	27,143	12,070	10,262
Insertion	18,983	4,066	2,293
Correct	46.06	76.01	79.61
Accuracy	8.33	67.93	75.05

Table 2. Estimation results based on the HMM.

(a) context-independent model

	Margin		
	0	1	2
Hit	8,806	28,214	38,847
Deletion	41,512	22,104	11,471
Insertion	35,372	16,253	5,915
Correct	17.50	56.07	77.20
Accuracy	-52.80	23.77	65.45

(b) context-dependent model

	Margin		
	0	1	2
Hit	14,198	35,967	42,611
Deletion	36,120	14,351	7,707
Insertion	32,970	11,521	5,110
Correct	28.22	71.47	84.68
Accuracy	-37.31	48.58	74.53

Table 3. Comparison of estimation performance between BRNN and MLP (Accuracy %).

Structure	Weights	Margin		
		0	1	2
MLP-1 (1 frame)	2,241	1.46	61.90	68.64
MLP-3 (3 frames)	2,241	6.12	64.16	70.94
MLP-5 (5 frames)	2,245	6.20	64.69	71.64
BRNN	2,181	8.33	67.93	75.05

Table 4. Estimation performance for the three kinds of BRNN based methods.

	Method		
	1	2	3
Hit	40,056	48,856	48,856
Deletion	10,262	1,462	1,461
Insertion	2,293	67,570	30,629
Correct	79.61	97.10	97.10
Accuracy	75.05	-37.19	36.22

described in **4.2.**. Here, method 3 would be applicable because 97.10% of the correct boundaries remain in the results with smaller insertion errors compared to those of method 2.

4. APPLICATION TO SPEECH RECOGNITION

From the experimental results, we can expect that the BRNN based segment boundary estimator gives useful information for existing speech recognition systems. In this section, we apply the BRNN based segment boundary estimator to two kinds of speech recognition systems, an HMM based system and a segment model based system, in order to achieve better recognition performance or reduce search space. For both systems, we investigate the effectiveness of phoneme boundary information through phoneme recognition experiments using the TIMIT database (462 speakers for training and 168 speakers for testing).

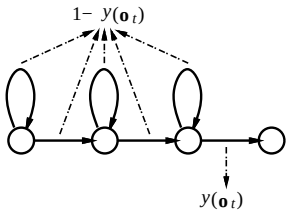

Figure 4. Integration of phoneme boundary information into an HMM.

4.1. HMM Based System

4.1.1. Integration as phoneme boundary probability

If we consider the outputs $y(\mathbf{o}_t)$ $(0 \leq y(\mathbf{o}_t) \leq 1)$ of the BRNN for the observation vector $\mathbf{o}_t$ at time t as the *phoneme boundary probability*, $1-y(\mathbf{o}_t)$ can be regarded as the *intra-phoneme probability*. These probabilities can be easily incorporated into the transition probabilities of conventional HMMs by taking the product of the two terms to be:

$$\bar{a}_{ij}(\mathbf{o}_t) = a_{ij} \cdot s_{ij}(\mathbf{o}_t), \tag{1}$$

where a_{ij} is the transition probabilities from state i to state j and

$$s_{ij}(\mathbf{o}_t) = \begin{cases} y(\mathbf{o}_t), & \text{if } j \text{ is the final state} \\ 1 - y(\mathbf{o}_t) & \text{else.} \end{cases} \tag{2}$$

Figure 4 shows an example of $s_{ij}(\mathbf{o}_t)$ for a three-state HMM. The time (observation) dependent transition probabilities $\bar{a}_{ij}$ are used in the recognition process.

4.1.2. HMM based recognition experiments

The shared-state context-dependent HMMs described in **3.1.** and a phoneme bigram language model were trained for 61 phonemes. Phoneme recognition was performed using a time-synchronous beam search based decoder [8]. Table 5 shows the recognition results and computational requirements compared to the total time of the utterances. We can see from this table that the BRNN-derived boundary probabilities improve not only recognition performance, but also computational requirements. Note that boundary probabilities can be obtained with small computational requirements: about 0.17 seconds are required for 3.06 second utterances on an HP735 workstation.

4.2. Segment Model Based System

4.2.1. Phoneme segment lattice creation

Recently, a variety of segment models (SMs) have been proposed for relaxing the independence assumption of observation, which is a shortcoming of conventional HMMs. SM based recognition systems, however, generally require much more computation than HMM systems. Therefore, to use SMs in in real-time systems, we have to reduce the computational costs by rescoring N-best candidates or word lattices obtained through HMM based recognition [9][10], or by generating segment lattices with a simple phoneme boundary detector [3]. However, it is not easy to improve performance unless accurate segment boundaries

Table 5. HMM based recognition results. Recognition results evaluated with 39 phoneme sets are shown in brackets.

	without language model		with language model	
	phoneme accuracy (%)	CPU time (%)	phoneme accuracy (%)	CPU time (%)
without boundary prob.	50.05 (58.69)	114.1	57.37 (64.71)	476.6
with boundary prob.	53.13 (61.75)	92.5	58.03 (65.47)	332.7
improvement (%)	6.2 (7.4)	18.9	1.5 (2.2)	30.2

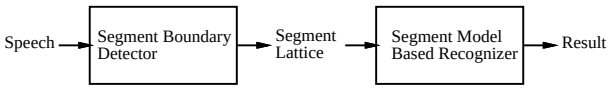

Figure 5. Block diagram of the BRNN-PSM based recognition system.

are included in the lattices. The fact that the segmentation probability gave a statistically significant improvement of recognition [5] indicates that accurate boundary estimation in a segment based recognition system is very important.

In method 3 described in **2.3.**, a boundary that is detected from the output at t and above threshold h is called a *main boundary*, and a boundary that is a local maximum between a lower threshold $l(< h)$ and h is called a *secondary boundary*. A segment lattice can be created by fully connecting boundaries existing between main boundaries. This lattice is used for phoneme recognition based on polynomial segment models (PSM) [11]. Figure 5 shows a block diagram of the BRNN-PSM based recognition system.

4.2.2. BRNN-PSM based recognition experiments

We generated a context-independent PSM with a single mixture for 61 phonemes. The regression order of the mean trajectories was set to 2. The variance was time invariant throughout a segment. The duration probabilities, which were computed from a histogram of the training segment durations, were used in the recognition. No language model was used in this experiment. Thresholds were set to $h = 0.6$ and $l = 0.25$.

Recognition results are listed in Table 6. For comparison, results obtained using three-state context-independent HMMs with a single mixture per state are also listed in the table. Table 7 shows the number of connections in the lattice (i.e. the number of segments to be evaluated) for all test data. "fully connected" indicates all possible connections with durations are from 3 frames to 70 frames. According to these results, the BRNN-PSM based method achieved both better recognition performance than the HMM system and a considerable computational reduction.

Note that the recognition performance of the BRNN-PSM based method will be improved by using a more precise PSM whose variant is time variance through a segment [11].

5. CONCLUSION

A segment boundary estimation method based on a BRNN has been proposed. The proposed method can accurately estimate segment boundaries by only using time series feature parameters. We applied this method to a speech recognition system and showed that (1) the usage of BRNN outputs was effective for improving the recognition rate and reducing computational time in an HMM based recognition system and (2) segment lattices obtained by the proposed methods dramatically reduce the computational complexity of segment model based recognition.

Table 6. Recognition results using BRNN based phoneme lattices and polynomial segment models (BRNN-PSM). Recognition results evaluated with 39 phoneme sets are shown in brackets.

	phoneme accuracy (%)
HMM	40.08 (49.64)
BRNN-PSM	41.80 (52.40)

Table 7. Computational reduction.

	# of segments	reduction rate
fully connected	2.46×10^7	—
proposed	73,348	1/335

ACKNOWLEDGMENT

The authors would like to thank Michiel Bacchiani of Boston University for his help involving our segment model based recognizer.

REFERENCES

[1] T. Svendsen and F. K. Soong : "On the automatic segmentation of speech signals," *Proc. ICASSP-87*, pp. 77–80, 1987.

[2] A. Ljolje and M. Riley : "Automatic segmentation and labeling of speech," *Proc. ICASSP-91*, pp. 473–476, 1991.

[3] J. Glass, J. Chang and M. McCandless : "A probabilistic framework for feature-based speech recognition," *Proc. ICSLP-96*, pp. 2277–2280, 1996.

[4] S.-L. Wu, M. Shire, S. Greenberg and N. Morgan : "Integrating syllable boundary information into speech recognition," *Proc. ICASSP-97*, pp. 987–990, 1997.

[5] J. Verhasselt, I. Illina, J.-P. Martens, Y. Gong and J.-P. Haton : "The importance of segmentation probability in segment based speech recognizers," *Proc. ICASSP-97*, pp. 1407–1410, 1997.

[6] M. Schuster : "Learning out of time series with an extended recurrent neural network," Neural Network Workshop for Signal Processing 96, pp. 170–179, 1996.

[7] J. Takami and S. Sagayama : "A Successive State Splitting Algorithm for Efficient Allophone Modeling," *Proc. ICASSP-92*, pp. 573–576, 1992.

[8] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga and Y. Sagisaka : "Spontaneous dialogue speech recognition using cross-word context constrained word graphs," *Proc. ICASSP-96*, pp. 145–148, 1996.

[9] H. Gish and K. Ng: "A Segmental Speech Model with Applications to Word Spotting," *Proc. ICASSP-93*, pp. II-447–II-450, 1993.

[10] A. Kannan and M. Ostendorf : "A comparison of constrained trajectory models for large vocabulary speech recognition," Boston Univ. Electrical and Computer Engineering Tech. Report ECE-96-007, 1996.

[11] T. Fukada, Y. Sagisaka and K. K. Paliwal : "Model parameter estimation for mixture density polynomial segment models," *Proc. ICASSP-97*, pp. 1403–1406, 1997.