SYLLABIFICATION OF CONVERSATIONAL SPEECH USING BIDIRECTIONAL LONG-SHORT-TERM MEMORY NEURAL NETWORKS

Christian Landsiedel[1,2]*, Jens Edlund*[1]*, Florian Eyben*[2]*, Daniel Neiberg*[1]*, Björn Schuller*[2]

[1]Department for Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden
[2]Institute for Human-Machine Communication, Technische Universität München, Germany
[1]{lastname}@speech.kth.se, [2]{lastname}@tum.de

ABSTRACT

Segmentation of speech signals is a crucial task in many types of speech analysis. We present a novel approach at segmentation on a syllable level, using a Bidirectional Long-Short-Term Memory Neural Network. It performs estimation of syllable nucleus positions based on regression of perceptually motivated input features to a smooth target function. Peak selection is performed to attain valid nuclei positions. Performance of the model is evaluated on the levels of both syllables and the vowel segments making up the syllable nuclei. The general applicability of the approach is illustrated by good results for two common databases—Switchboard and TIMIT—for both read and spontaneous speech, and a favourable comparison with other published results.

Index Terms— Syllabification, Recurrent Neural Networks, Speech Analysis, Dialogue Systems

1. INTRODUCTION

Temporal structure is an important source of information describing speech. Between the fine-grained phone and the coarser word and utterance levels, syllabic segmentation provides insight both in the phonological and rhythmic aspects of speech. It can be put to use in various applications, for example in analysing dialogue for human-machine interaction [1]. The identification of the vowel constituting the core of a syllable, the *syllable nucleus*, is a key part in measuring speech prominence [2], and speech recognition has been helped by information about the syllabic structure of utterances [3], such as the number of syllables in an utterance or the position of syllable nuclei.

We present a novel, data-driven syllabification approach based on Long-Short-Term Memory (LSTM) Neural Networks [4, 5]. It is primarily aimed at the use as a tool for dialogue analysis and spoken dialogue systems. Its general applicability is demonstrated by experiments on two well-known databases—Switchboard and TIMIT—of both read and spontaneous speech.

The remainder of this paper is structured as follows: We first introduce relevant known approaches to syllabification in Section 2, then our suggested novel approach employing long-term memory enhanced recurrent neural networks in Section 3, before presenting our experiments and results—including a comparison of the obtained outcomes with other published results—in Section 4 and concluding in Section 5.

2. APPROACHES TO SYLLABIFICATION

2.1. Contour-based Approaches

A number of syllabic segmentation strategies rely on the extraction of a one-dimensional envelope from the acoustical signal. A peak search procedure on this contour provides a set of local minima and maxima, serving as estimates for syllable nucleus or syllable border positions, respectively.

One example for this class of approaches is the calculation of an energy envelope by low-pass filtering the signal energy in a broad frequency range [6]. Segmentation is then carried out with a recursive *convex hull* algorithm, which exploits thresholds on the relationships between adjacent maxima and minima of the envelope to find syllable borders and nuclei. This approach is expanded on by using an altered frequency range for the calculation of the energy envelope and performing peak selection with a multilayer perceptron evaluating peak intensity as well as segment duration [7].

The approach described in [8] mainly relies on extracting a spectral correlation envelope from selected subband energies, in addition to temporal correlation and smoothing. Peak picking is performed on this envelope employing duration and value thresholds as well as a voicing decision.

Rhythm guidance has been proposed in order to incorporate a modelling of speech rhythmicity into syllable nucleus detection [9]. An instantaneous speech rhythm is iteratively estimated by fitting a sinusoid function to all positions in the set of hitherto estimated peaks in the utterance. The length of a search interval following the last estimated nucleus is then determined as a multiple of the estimated sinusoid's period length, and the local maximum with the highest value inside that interval is used as estimate for the next nucleus position. Estimates in unvoiced segments are discarded.

2.2. Machine Learning Approaches

Alternative to the combination of the calculation of a signal contour and a peak-picking scheme, a different class of algorithms is built on modelling syllabicity statistically with machine learning methods. As an example for these data-driven methods, the use of a multilayer perceptron for the task of syllable onset detection has been evaluated [10].

As a different learning-based method, a standard recurrent neural network model and a recurrent *Temporal Flow Model* architecture allowing various delays for the connections between different network layers are compared for the task of finding syllable borders [11].

This research was carried out at KTH Speech Music and Hearing, Stockholm, supported by the Swedish Research Council project *Rhythm of conversation* (VR2009-1766).

At the time of writing the last author was an invited researcher at the Imperial College London's Department of Computing in London/UK.

3. SYLLABIFICATION WITH A BLSTM NETWORK

3.1. Bidirectional Long-Short Term Memory Networks

The concept of Long-Short Term Memory (LSTM) [4, 5] increases the availability of temporal contextual information for a neural network. This is achieved by replacing the nodes of a standard recurrent network by interconnected blocks, in each of which information is stored in one or more internal state cells. Input and output of new data as well as decay of the internal state recurrence are controlled multiplicatively by gate nodes, so that the memory blocks to some extent resemble the memory cells used for binary storage in semiconductor memory. A single LSTM block showing the architecture of the internal state node and the gate nodes is depicted in Figure 1. With this structure, the way information is stored and used at different time steps can be controlled much more flexibly than in traditional neural networks, since the behaviour of the controlling gates is learned during the training process of the network. The added increase in temporal modelling promises improvements for segmentation problems such as the syllabification task at hand, since long-term temporal structure existing in the data can be learned automatically.

A bidirectional architecture gives the network access to past and future contextual information, requiring the whole information to be present at the beginning of the processing. For the syllabification method developed here, a bidirectional architecture with two hidden layers and 30 LSTM cells in each was trained using the on-line gradient descent algorithm at a learning rate of 10^{-4} with a momentum of 0.9. 8-fold cross validation was used in combination with *early stopping* [5], so that the network training was stopped when no performance increase on a disjunct validation set had been registered for a run of 20 epochs. The models at the epoch with the best validation set performance were used for evaluation.

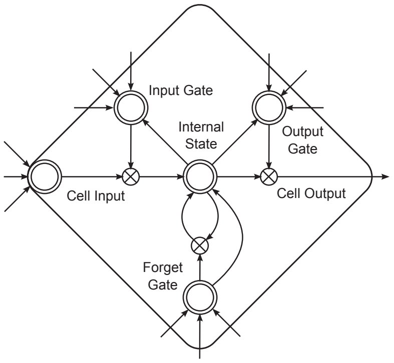

Fig. 1: An LSTM block with a single internal state cell

3.2. Data Representation

Different spectral representations were considered for the use as input representation for the BLSTM network. Since a combination of temporally smoothed and non-smoothed representations as well as the inclusion of delta features proved beneficial, a concatenation of a 20-band modulation spectrum [12] calculated on a critical-band-warped frequency scale and their first differences and an ensemble consisting of 12 PLP coefficients [13] and logarithmized energy and their first and second differences were used as features. Feature extraction was partly done using the openSMILE utility [14].

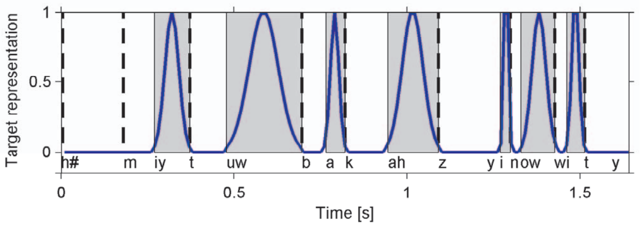

Fig. 2: Target representation for an example utterance. Vowel segments are shaded, syllable borders are indicated by broken vertical lines.

In order to specify a desired output shape, Gaussian segments spanning target segments from the annotation were concatenated with constant zero-valued stretches for non-vocalic segments, as illustrated in Figure 2. The neural network was trained to perform non-linear regression from the 79-dimensional input vector sequence to the thus specified target function, thereby incorporating both information about syllable nucleus position via its peak structure, and the temporal extension of the vowels.

3.3. Peak Picking and Threshold Calculation

The output of the BLSTM network is the activation of the output node for each frame presented to the network in the input sequence. An example of an output sequence is shown in Figure 3.

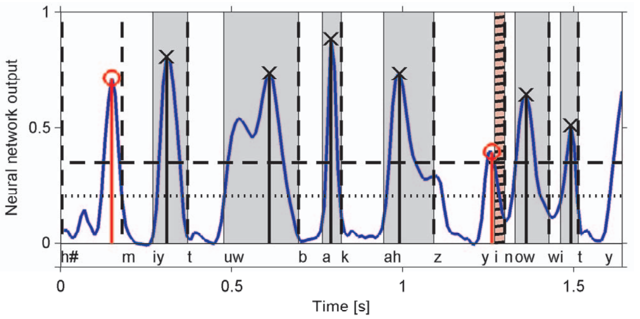

Fig. 3: Output of the BLSTM network with valid peaks according to region-based thresholding for an example segment from the Switchboard corpus. Vowel regions identified correctly are shaded grey, missed ones are hatched; syllables are delineated by vertical broken lines. Correctly placed nuclei are denoted by crosses, inserted ones by circles. The minimum separation threshold T_l (dashed) used is the segment output mean value (dotted) multiplied with a factor determined for each cross validation configuration on the validation set.

While most short vowels are present in the neural network's output as clearly defined, narrow peaks, multiple peaks may occur for lengthened vowels, such as the back-channels occurring frequently in spontaneous dialogue. However, these peaks are commonly not separated by dips in the neuronal network's output values as deep as those occurring between separate vowels. Therefore, a region-based peak selection strategy was adopted. Candidate regions are identified as those regions in which the neuronal network's output is continuously above a lower threshold T_l, and the position of the maximum within each candidate region is taken as a nucleus estimate. The minimum separation threshold T_l is estimated as a multiple of

Table 1: Description of the data sets used. Each set was split in 8 equal-sized subsets for cross validation.

	# turns	min/max length [s]	# vowels/syllables
STP	6 40	0.45 / 13.73	5 813 / 5 806
TIMIT	1 344	1.06 / 7.44	17 283 / 17 283

the mean segment output value, where the factor is determined for each cross validation configuration so as to minimize the total error on the corresponding validation set.

In order to analyse the effect of additional temporal modelling on top of this method, an explicit rhythm-guided peak selection explicitly modelling the distance between syllable nuclei was added. To this end, a context of a maximum of 4 subsequent internuclei distances was modelled with a 4-component Gaussian Mixture Model for the logarithms of the distances. To select valid peaks from the set of candidates provided by the region-based thresholding scheme, each utterance was traversed from left to right. In each step, the most probable candidate was selected based on the probability distribution provided by the Gaussian model conditioned on the distances between the 3 preceding peaks. The parameters of the model were estimated from the annotation of the utterances used for training for each cross validation configuration.

4. EXPERIMENTS AND RESULTS

4.1. Data Sets

The BLSTM syllable nucleus estimation was trained and evaluated both on read and spontaneous speech data. In an effort to find a trade-off between training time and coverage of the data, a subset of each corpus was used for training and evaluation of a respective model as described below. An overview over the subsets from each corpus used is given in Table 1. Conversational speech data was taken from the subset of the Switchboard Database annotated on a syllabic level by the Switchboard Transcription Project (STP) [15]. In order to have a valid representation of the syllabic vowels, abutting vowel segments that form a diphtong within a syllable were combined, and pauses filled with vowelic sounds were added as target segments for syllabic nuclei.

An additional model was trained and evaluated on a set made up of 1 344 utterances from the TIMIT corpus of read speech. In the selection of the sentences, care was taken that no single speaker or single sentence occurred in more than one group for cross-validation. The dialectal (*sa*) sentences were excluded for their similar syllabic structure. Based on the available phoneme annotation for the Switchboard corpus, syllabic annotation was created using the rule-based syllabification implemented in the *tsylb2*[1] program.

4.2. Evaluation Procedure

For the evaluation of the model performance, either the phonetic and syllabic segmentations stemming from the annotation of the corpora used were taken as the ground truth. The scoring procedure counted one-to-one matches between a nucleus estimate and a target segment as correct, target segments missed entirely as deletions and surplus nucleus estimates in target segments as well as estimates in non-target regions as insertions. Thus, segments containing more than one estimate were counted neither as correct nor as deletions, but all but one estimates in this segment were evaluated as insertions, which accounts for the differences between the rate of correct estimations and the recall values. Alternative evaluation schemes were adopted in order to be able to compare performance with published results.

4.3. Results

Using this evaluation method, the BLSTM syllabification models achieved the results reported in Table 2 on the used datasets from TIMIT and the STP-annotated Switchboard corpus. Figures averaged over the performance of all models generated in an 8-fold cross-validation training run with early stopping as described in section 3.1. Results are given both for annotated vowel and syllable segments, and for the region-based and the rhythm-guided peak picking methods.

Table 2: Results of the BLSTM approach for phone and syllable level evaluation as described in section 4.2 with a minimum separation threshold (MST) peak selection procedure and added rhythm-guidance (RG). The results are averaged over those generated by 8-fold cross-validation.

(a) ICSI-annotated Switchboard data

STP [%]	phone level		syllable level	
	MST	MST+RG	MST	MST+RG
correct	**82.29**	81.22	**82.65**	82.03
insertion	16.77	**15.05**	16.28	**14.11**
deletion	**15.18**	16.78	**11.93**	13.45
recall	**84.44**	82.88	**87.40**	85.92
precision	83.11	**84.40**	83.58	**85.34**
F_1	**83.74**	83.61	85.42	**85.61**

(b) TIMIT data set

TIMIT [%]	phone level		syllable level	
	MST	MST+RG	MST	MST+RG
correct	**90.53**	88.10	**90.25**	88.43
insertion	5.72	**4.71**	5.99	**4.38**
deletion	**8.22**	11.00	**6.88**	9.49
recall	**91.68**	88.90	**92.92**	90.31
precision	94.07	**94.94**	93.79	**95.29**
F_1	**92.85**	91.81	**93.34**	92.73

For comparison, these results can be related to the ones published in [8] for the algorithm based on temporal correlation and spectral cross-correlation introduced in section 2.1. It is mainly evaluated on a syllable rate correlation measure, but figures for an evaluation in terms of a one-to-one mapping of peaks in the correlation envelope and the syllabic transcription are also given. Note, however, that insertions are based on a count of segments in this evaluation, not on a count of surplus peaks within a target segment or outside a target region. In Table 3, the results published in [8] are given and compared with a similar evaluation of the BLSTM approach with region-based peak search in Table 3.

In Table 4, performance of the BLSTM approach with region-based peak search on the TIMIT set is compared to results published in [9]. Here, the evaluation method is based on peaks coinciding with annotated vowel segments, which are padded symmetrically to a minimum length of 50 ms if necessary.

The above comparisons show that the introduced novel BLSTM approach compares equally or favourably with other methods: The observed superiority in F_1 measure in Tables 3 and 4 is significant at the 0.05 level for STP and at level $\ll 10^{-3}$ for TIMIT in a one-tailed

[1] written by Bill Fisher, no citation available. The program is available at ftp://jaguar.ncsl.nist.gov/pub/, last acc. Oct 12, 2010

Table 3: Comparison of the performance of the temporal and spectral cross-correlation (TCSSC) approach, as given in [8] and the introduced BLSTM approach on Switchboard data, evaluated on a syllable-segment-based evaluation, where syllable segments with multiple peaks are counted as a single insertion.

STP [%]	**TCSSC** [8]	**BLSTM**
correct	80.60	**82.65**
insertions	**3.80**	5.42
deletions	15.60	**11.93**
recall	83.78	**87.39**
precision	**95.50**	93.84
F_1	89.26	**90.50**

Table 4: Comparison of performance on TIMIT data of the introduced BLSTM approach to the methods evaluated in [9]: an implementation of [8] (TCSSC), an energy-based baseline method (nRG) and rhythm-guided syllabification [9] (RG). Evaluation is done on the basis of syllable nuclei estimates coinciding with annotated vowel segments, where short segments are symmetrically padded to a minimum length of 50 ms if necessary.

TIMIT [%]	**TCSSC** [9]	**nRG** [9]	**RG** [9]	**BLSTM**
recall	86.06	79.97	86.59	**92.22**
precision	99.69	**99.84**	98.86	95.82
F_1	90.21	88.58	92.07	**93.98**

test. At the same time no further parametrization apart from the design choices of the network architecture and the input representation is needed. Thereby, as there is no need for heuristics and the specific formulation of knowledge about the nature of the problem, the transfer of the approach to other domains and languages is simplified.

The temporal precision of the approach can be seen by the fact that results evaluated on syllable and phoneme level are similar, showing that the majority of estimates are within the salient part of the syllable and can serve both as vowel and syllable nucleus landmarks. The fact that the direct inclusion of inter-nuclei duration modelling in the rhythm-guided peak selection did not lead to improvements over the method underlying it shows that also these long-term temporal relationships are represented in the BLSTM approach.

The complex nature of the conversational speech in the Switchboard data set poses a greater problem for syllabification, as can be seen by the considerable difference to the results for TIMIT. Part of this difference can also be explained with the problems arising at the definitions of target segments for training and evaluation for those parts of the data which are non-lexical. The training of data-based models for segmentation and all evaluation depend crucially on data annotation, which is not standardized for the frequent non-lexical segments in spontaneous speech. Therefore, for the application in dialogue systems and for speech rate evaluation, other, perceptually motivated segmentation levels should be evaluated in addition to the syllabic level. Since we believe the BLSTM approach to be more generally applicable for a wider class of speech segmentation problems, further work can be done on extending this method to those domains.

5. CONCLUSION

A new approach to the problem of syllable nucleus detection based on the BLSTM neural network formulation has been presented. Its applicability to the task of syllabification of both read and spontaneous speech has been shown by using it on data with very diverse characteristics. Models both for TIMIT data and for data from the Switchboard corpus showed good results, without the need of explicitly specifying different parametrizations for the different corpora.

The syllabification model introduced here will be used in experiments to further the understanding of temporal aspects of speech in dialogue. Coming experiments will include the analysis of the behaviour of dialogue partners on data from the full Switchboard dialogues and from the SPONTAL database [16] of Swedish multi-partner conversations.

6. REFERENCES

[1] N. Fujiwara, T. Itoh, and K. Araki, "Analysis of changes in dialogue rhythm due to dialogue acts in task-oriented dialogues," in *Proc. 10th Intl. Conf. Text, Speech and Dialogue*, 2007, pp. 564–73.

[2] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 15, no. 2, pp. 690–701, 2007.

[3] C. D. Bartels and J. A. Bilmes, "Use of syllable nuclei locations to improve ASR," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Kyoto*, 2007.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–80, 1997.

[5] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, Ph.D. thesis, Technische Universität München, München, July 2008.

[6] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. Am*, vol. 58, no. 4, pp. 880–3, 1975.

[7] A.W. Howitt, *Automatic syllable detection for vowel landmarks*, Ph.D. thesis, MIT, 2000.

[8] D. Wang and S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 15, no. 8, pp. 2190–201, nov. 2007.

[9] Y. Zhang and J. R. Glass, "Speech rhythm guided syllable nuclei detection," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2009, pp. 3797–800.

[10] S.-L. Wu, M.L. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1997, vol. 2, pp. 987–90.

[11] L. Shastri, S. Chang, and S. Greenberg, "Syllable detection and segmentation using temporal flow neural networks," in *Proc. 14th Int. Congr. Phonetic Sciences*, 1999, pp. 1721–24.

[12] B.E.D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 1-3, pp. 117–32, 1998.

[13] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–52, 1990.

[14] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (MM)*, Firenze, Italy, 2010, ACM.

[15] S. Greenberg, "The Switchboard transcription project," in *LVCSR Summer Workshop Technical Reports*, 1996.

[16] J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, and D. House, "Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture," in *Proc. 7th Int. Conf. Language Resources and Evaluation*, 2010, pp. 2992–5.