SYLLABLE DETECTION AND SEGMENTATION USING TEMPORAL FLOW NEURAL NETWORKS

Lokendra Shastri, Shuangyu Chang, Steven Greenberg
International Computer Science Institute
{shastri, shawnc, steveng}@icsi.berkeley.edu

ABSTRACT

The syllable serves as an important interface between the lower-level (phonetic and phonological) and the higher-level (morphological and lexical) representational tiers of language. It has been demonstrated that reliable segmentation of spontaneous speech into syllabic entities is useful for speech recognition. An automatic method is described for delineating the temporal boundaries of syllabic units in continuous speech using a Temporal Flow Model (TFM) and modulation-filtered spectral features. The TFM is a neural network architecture that supports arbitrary connectivity across layers, provides for feed-forward as well as recurrent links, and allows variable propagation delays along links. Two TFM configurations, global and tonotopic, have been developed and trained on a phonetically transcribed corpus of telephone and address numbers spoken over the telephone by several hundred individuals of variable dialect, age and gender. The networks reliably detected the boundaries of syllabic entities with an accuracy of ca. 84%.

1. INTRODUCTION

The syllable is playing an increasingly important role in the design of automatic speech recognition (ASR) systems, providing an intermediate representation capable of binding the lower-level (phonetic and phonological) linguistic tiers with those germane to the lexicon and grammar [3]. The syllable's significance makes it a useful representational unit for developing future-generation speech recognition systems capable of reliably detecting and segmenting syllabic entities in the acoustic speech signal.

Most current-generation ASR systems for English use the phonetic segment as the basis from which to derive lexical information from the acoustic signal. Although this phone-based approach has been moderately successful for carefully enunciated speech under pristine acoustic conditions, it has been less useful for recognizing such material under real-world conditions (i.e., background noise and reverberation) or when the speech is of a form more characteristic of informal conversation [7].

One potential reason why current ASR systems do so poorly under such conditions is that lexical units are represented solely as sequences of phonetic segments. Because automatic segmentation and labeling of speech at the phonetic-segment level (a.k.a. automatic alignment) are not very accurate (compared to the segmentation and labeling performed by trained phoneticians [4]), ASR word models are inherently fragile and often "break" under a wide variety of environmental and linguistic conditions. Increasing the stability of lexical models is likely to result in significant gains in speech recognition performance [8] and there are a number of reasons why stable lexical models are more readily based on syllables than on phones. Primary among these are the close statistical association between words and syllables (in English) and the structural integrity of the syllabic onset, nucleus and coda [3]. Moreover, it has recently been shown that an ASR system based on syllabic units is more accurate (when combined with the classic phone-based approach) than a system based purely on phonetic segments [10][15][16]. In view of the above, we believe that segmentation of the acoustic signal into syllabic segments is an important stage in the development of a syllable-centric ASR system.

2. TEMPORAL FLOW MODEL

We perform syllabic segmentation using a neural network architecture based on the Temporal Flow Model (TFM) of Watrous and Shastri [13]. TFM supports arbitrary link connectivity across multiple layers of nodes, admits feedforward as well as recurrent links, and allows variable propagation delays to be associated with links (cf. Figures 1 and 2). The recurrent links in TFM provide a means for smoothing and differentiating signals, measuring the duration of features, and detecting their onset. The use of multiple links with variable delays allows the system to maintain context over a window of time and thereby carry out spatio-temporal feature detection and pattern matching. In combination, the use of recurrent links and variable propagation delays provide a rich mechanism for simulating such properties as short-term memory, integration and context sensitivity — properties that are essential for processing time-varying signals. In the past TFM has been successfully applied to a number of phoneme-recognition tasks covering a broad range of the articulatory space [11][14] as well as hand-printed digit recognition [9].

3. TELEPHONE DIGIT RECOGNITION TASK

In order to explore the feasibility of applying TFM to syllable segmentation we have limited its current application to the domain of numerical sequences spoken over the telephone. As our experimental test-bed we chose a subset of the Numbers95 corpus [2] containing "fluent" numbers such as are spoken in the context of household addresses.

Each word in this corpus has been labeled and segmented at the phonetic-segment level by a linguistically trained individual and these materials have been automatically syllabified. Numbers95 contains only 33 separate syllables, making the corpus particularly useful for the development of novel recognition algorithms. Despite the restricted size of the lexicon, the corpus contains speech spoken by a large number of individuals (of both genders) spanning a wide range of geographical dialects, speaking rates and variable utterance length.

4. MODULATION SPECTROGRAM

Prior to syllabic segmentation the acoustic signal is transformed into a modulation-filtered spectrogram (MSG) [5][8]. This representation encodes the speech signal in terms of low-frequency energy ($<$ 16 Hz) across time and frequency. The statistical properties of the modulation spectrum have been shown to closely reflect the duration of syllabic segments in spontaneous speech, with a peak in the distribution at ca. 5 Hz [3][4]. Significant alteration of the modulation spectrum has a deleterious effect on speech intelligibility [3].

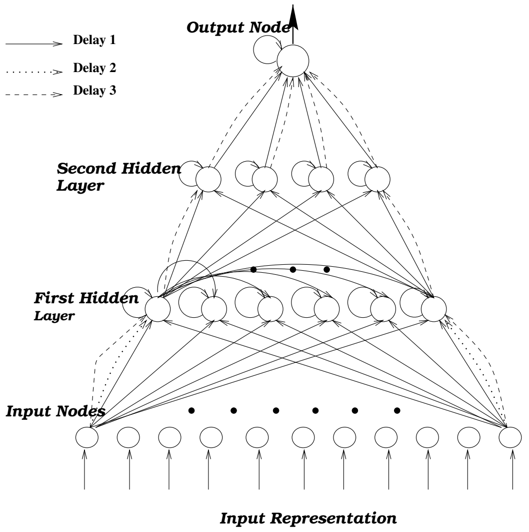

Figure 1: Architecture of the global Temporal Flow Model neural network used for syllabic segmentation.

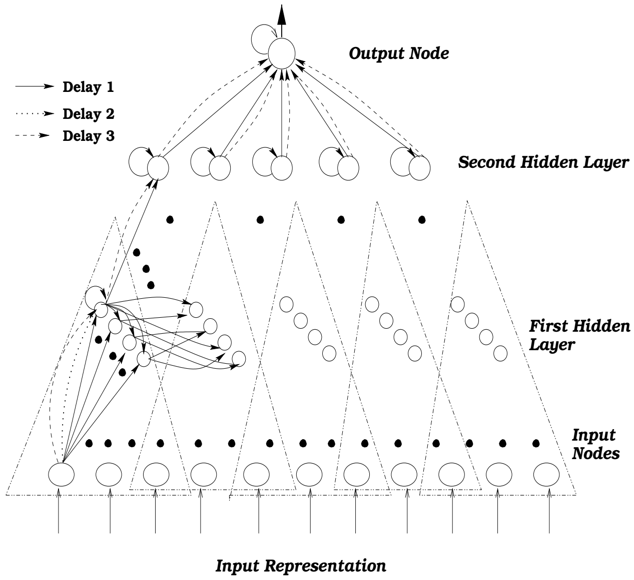

Figure 2: Architecture of the tonotopic Temporal Flow Model neural network.

For the current study the spectrum was partitioned into 11 discrete, critical-band-like (1/4-octave) channels, over which the modulation spectrogram was computed using a 250-ms, Hamming window using a slide interval of 10-ms.

5. TASK FORMULATION

The immediate goal was to develop a TFM capable of automatically identifying the onset and offset of syllabic constituents using the MSG representation of the speech signal.

The target outputs for training the neural networks were derived from the syllable-level transcription of the utterances. The duration of each syllable was ascertained from the segmentation information in the transcription and a Gaussian curve plotted over the syllabic segment as the target output. The target outputs were then normalized across all syllables and scaled so that the height of the Gaussian curves were directly proportional to the length of the associated syllables. The Gaussian target outputs were intended to provide only a rough approximation to the network during training as a means of obtaining the sort of response desired from the network.

6. NETWORK MODELS

Two distinct TFM network configurations were investigated, one with global connectivity (Figure 1), the other with tonotopic connectivity (Figure 2). Both configurations contained an input layer, two hidden layers (H1 and H2), and an output layer. The input layer in both configurations contained eleven nodes - one for each of the eleven MSG features. The two network configurations differed, however, in (a) how the input layer was connected to H1 and (b) the density of lateral connections within H1. In the global configuration, all input nodes were connected to all H1 nodes and all H1 nodes were densely connected via lateral links. In the tonotopic configuration, H1 nodes were divided into distinct groups, each receiving activation from a small number of adjacent input nodes (i.e., channels). Nodes within a group were densely connected, but nodes across groups had only sparse interconnections. In both configurations, H1 nodes were fully connected to H2 nodes which, in turn, projected to the output node.

Figure 1 shows a typical configuration of the global model. The model has 11 input nodes, each receiving an MSG feature. H1 and H2 consist of hidden nodes with self-recurrent links. Between each input node and each node in H1 there are three separate links, each with a different propagation delay (1, 2, 3). Nodes within H1 are also connected with lateral links. Between each node in H1 and each node in H2, there are two links with delays 1 and 3, respectively. Nodes in H2 are connected to the output node via a similar constellation of links. In general, the number of links, propagation delays and the number of hidden nodes can vary depending on the task.

Figure 2 shows a typical configuration of the tonotopic model. The hidden nodes in H1 are divided into five distinct groups. Each of these receives activation from three adjacent input nodes. The input nodes of adjacent groups overlap by a factor of one (i.e., the "receptive fields" of two adjacent groups overlap by 1). An H1 node receives three links with propagation delays of 1, 2 and 3, respectively, from each input node in its receptive field. All nodes within a group are fully connected with links of different propagation delays. Nodes across groups are also connected via links of different propagation delays, but these links are quite sparse. The H2 nodes receive two links from each H1 node with propagation delays of 1 and 2, respectively. H2 nodes are also fully linked to the output node in a similar manner. In general, the size of, and the overlap between, the receptive fields of H1 nodes, the number of nodes within each group in H1, the number of links, propagation delays and the number of H2 nodes can vary depending on the task.

7. TRAINING PROCEDURES

The specification and training of the neural networks was performed with GRADSIM [12], a connectionist network simulator that uses gradient optimization techniques. GRADSIM allows different delays to be associated with links, it supports fixed and modifiable link-weights, it admits feedforward as well as recurrent architectures, and it implements a number of different optimization methods. For this specific application we used the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [1][12], a second-order, gradient-based optimization algorithm because of its speed and accuracy. The network was trained on a subset of training-set sentences. At each iteration the trained parameters were tested on a smaller cross-validation set. To avoid overfitting, the training was stopped when the error rate on the cross-

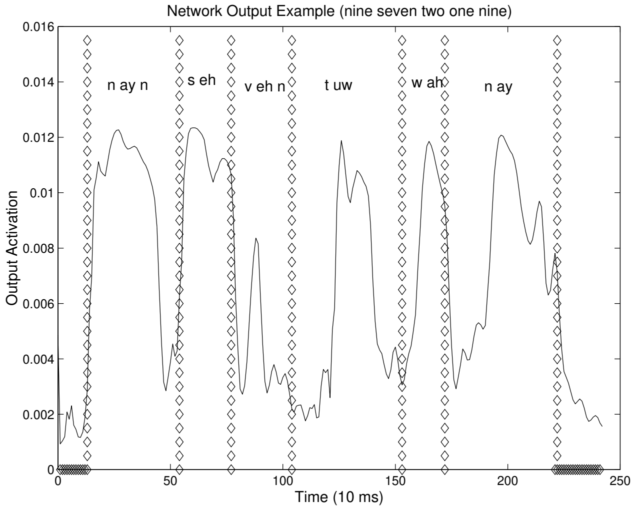

Figure 3: Sample output of a TFM network (tonotopic model). The vertical lines (composed of diamonds) are actual syllable boundaries. The heavy dots on the time axis indicate segments of silence.

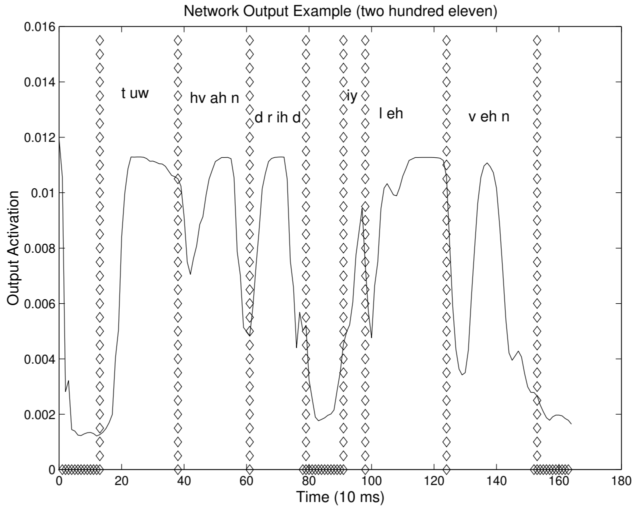

Figure 4: Sample outputs of a TFM network (global model).The vertical lines (composed of diamonds) are actual syllable boundaries. The heavy dots on the time axis indicate segments of silence.

validation set stopped decreasing for a pre-specified number of iterations. Network output samples at the convergence point are shown in Figures 3 and 4.

7.1. APPLICATION TO SYLLABLE DETECTION AND SEGMENTATION

The accuracy of the network output can be evaluated in a preliminary fashion by visual inspection. However, to provide a quantifiable metric for comparison of different networks, we applied the trained network to predict the presence of syllables in sentences for which it had never been trained. Automatic procedures were created to convert the network outputs to syllable detection, and eventually, accuracy scores in terms of percentages of false negative and false positive responses. Before applying the conversion procedure the network outputs were processed with a simple low-pass filter to obtain a smoother representation of the outputs. The conversion procedure and the filtering only depend on a limited number of frames (usually no more than three frames) around the current frame of interest. The entire system can be implemented in an on-line form with minimal delay.

7.1.1. Two-level Threshold Syllable Detection The simplest method for syllable detection is to statically set a threshold for the network outputs. A syllable onset is detected whenever the network output crosses the threshold in an upward-going trajectory. The temporal position of the detected syllable is compared with that associated with the boundary derived from manual segmentation. If the onset of the automatic system is within a certain tolerance limit of the human-delimited boundary then the automatically defined onset is considered to have been accurately determined (a "hit"). If the network's onset lies outside of the tolerance window the event is scored as a "miss." The threshold level used can be determined empirically on a validation data-set. The drawback of using a single, fixed threshold is that it ignores the variations in magnitude of the network output associated with the duration and amplitude of the syllabic sequences.

A two-level dynamic thresholding method was developed to minimize such problems. Thus, in addition to a fixed higher-level threshold, a dynamic lower-level threshold was also used. Such a two-tiered threshold enabled the network to respond to syllables with weaker outputs levels. The distance between the higher- and lower-level thresholds was determined dynamically with a decaying, running average of the network output levels computed over the course of the utterance. Figure 5 illustrates this method in operation. The two-level thresholding reduced false-negative errors by ca. 30%. False-positive errors were reduced by setting a lower bound on the allowable syllabic duration (doing so, however, introduced some false-negative errors).

The two-level dynamic thresholding method together with the minimum syllabic duration heuristic yielded a reasonable level of performance. Table 1 indicates the prediction error on test utterances using the minimum syllabic duration heuristic and either the static threshold or the two-level dynamic thresholding algorithm. The results are shown for both a global network and a tonotopic network.

8. DISCUSSION

The modulation spectrogram was used as input to the neural networks because of its close association with syllabic entities as well as its fidelity to certain functional properties of upper stations of the auditory pathway. However, it is also possible to use other representational forms, such as RASTA [6] in order to segment speech at the phonetic-segment level (as has been done in preliminary experiments).

In many situations, instead of a "yes" or "no" response, it would be desirable to ascertain the probability associated with the occurrence of a syllable. The current network architecture is amenable to such a probabilistic framework and could provide such an input to an ASR system.

The TFM neural network architecture offers several advantages over the traditional feedforward multi-layer perceptron (MLP) architecture. A TFM network treats time in a transparent manner — the input to a TFM network at time t is simply the input signal at time t. In particular, the input nodes of a TFM network are not replicated n times to realize a context window of size n. In a TFM network, the requisite temporal integration of the input signal occurs within the network as a result of converging activity arriving along recurrent links and links with varying delays. The architecture of the tonotopic TFM model is particularly well-suited for extracting temporally extended features within limited frequency channels (in layer H1) and subsequently integrating several such

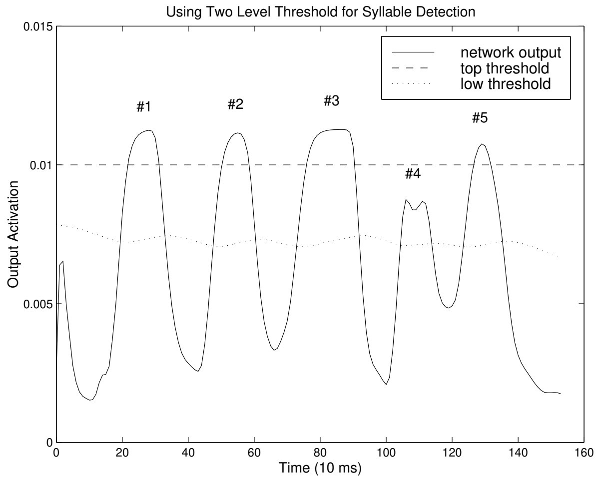

Figure 5: An illustration of the two-level threshold method for detecting syllables from network output. The network outputs have been low-pass filtered. Five syllables are detected in this example.

TFM Form	Threshold Levels	False Positives Percent	False Negatives Percent	Total Error Percent
Global	2	4.40	11.46	15.86
	1	4.50	14.17	18.67
Tonotopic	2	4.95	11.72	16.67
	1	6.60	13.53	20.13

Table 1: Experimental results of syllable detection using the minimum syllabic duration heuristic and either the one-level static threshold or the two-level dynamic thresholding algorithm. The training set consists of 100 sentences. The test set consists of 1000 sentences with a total of 5822 syllables. A false-positive response means that the network detected a syllable where none existed; a false-negative response means that the network failed to detect a syllable.

features across multiple frequency bands (in layer H2). The temporal extend of such features is determined within a network as a consequence of learning.

9. SUMMARY AND CONCLUSION

A Temporal Flow Model network has been developed to extract syllabic boundary information from continuous speech. The TFM naturally captures the time-varying properties of speech in a compact network representation. Two distinct forms of TFM networks have been applied to modulated-filtered spectrographic representations of the OGI Numbers corpus - a global model and a tonotopic model. The networks produce outputs incorporating syllabic information. This information can be used to predict the onset of syllabic entities with an accuracy of ca. 84%. Even for low-level tasks such as syllable onset detection, a "perfect" solution can often only be acquired by using higher-level grammatical and semantic knowledge. Ultimately, it would be useful to incorporate feedback from higher representational tiers steps into lower-level segmentation in order to enhance the system's performance.

ACKNOWLEDGMENTS

This research was supported by the Learning and Intelligent Systems program of the National Science Foundation.

REFERENCES

[1] Bishop, C. M. (1996) *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, England.

[2] Center for Spoken Language Understanding, Dept. of Computer Science and Engineering, Oregon Graduate Institute. (1995) Numbers corpus, Release 1.0.

[3] Greenberg, S. (1998) "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation," *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kekrade (Netherlands), pp. 47-56.

[4] Greenberg, S., Hollenback, J. and Ellis, D. (1996) "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," *Proceedings of the Fourth International Conference on Spoken Language*, S24-27.

[5] Greenberg, S. and Kingsbury, B. (1997) "The modulation spectrogram: In pursuit of an invariant representation of speech," *ICASSP-97, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1647- 1650.

[6] Hermansky, H. and Morgan, N. (1994) RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2, 578-589.

[7] Kingsbury, B. (1998) Perceptually Inspired Signal-Processing Strategies for Robust Speech Recognition in Reverberant Environments. Ph.D. Thesis, Univ. California, Berkeley.

[8] Kingsbury, B., Morgan, N. and Greenberg, S. (1998) "Robust speech recognition using the modulation spectrogram," *Speech Communication*, 25, 117-132.

[9] Shastri, L. and Fontaine, T. (1995) "Recognizing handwritten digit strings using modular spatio-temporal connectionist networks," *Connection Science*, 7(3,4), 211-245.

[10] Shire, M. L. (1997) "Syllable onset detection from acoustics," Master's Thesis, EECS Dept., University of California, Berkeley.

[11] Watrous, R. L. (1990) "Phoneme discrimination using connectionist networks," *Journal of Acoustic Society of America*, 87, 1753-1772.

[12] Watrous, R. L. (1993) "GRADSIM: a connectionist network simulator using gradient optimization techniques," Report, Siemens Corporate Research, Inc., Princeton, New Jersey.

[13] Watrous, R. L. and Shastri, L. (1986) "Learning phonetic features using connectionist networks: An experiment in speech recognition," Tech. Report, MS-CIS-86-78, University of Pennsylvania.

[14] Watrous, R. L. and Shastri, L. (1987) "Learning phonetic features using connectionist networks." *Proceedings of IJCAI-87, the Tenth International Joint Conference on Artificial Intelligence*, Milan, Italy, August 1987. pp. 851-854.

[15] Wu, S.-L., Kingsbury, B., Morgan, N. and Greenberg, S. (1998) "Incorporating information from syllable-length time scales into automatic speech recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 721-724.

[16] Wu, S.-L., Kingsbury, B., Morgan, N. and Greenberg, S. (1998) Performance improvements through combining phone- and syllable-length information in automatic speech recognition, *Proceedings of the International Conference on Spoken Language Processing,* pp. 854-857.