ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

Duration Modeling For Telugu Language with Recurrent Neural Network

V.S.Ramesh Bonda, P.N.Girija

Professor, School of Computer & Information Sciences, University of Hyderabad, Hyderabad, Andhra Pradesh, India

ABSTRACT: In this paper, a novel syllable duration modeling approach for Telugu speech is proposed. Duration of a syllable is influenced by positional and contextual variations of syllables. Multiple linguistic features of syllables at different levels like positional and contextual features are used from text. Duration values of syllables are extracted from speech analysis software PRAAT. Duration of a syllable is predicted by a Recurrent Neural Network (RNN) algorithm. A small speech database is considered as a preliminary work to predict syllable duration with proposed RNN algorithm. Experiments are conducted with different sets of features.

KEYWORDS: duration, speech synthesis, recurrent neural networks, syllables, Parts of Speech, Positional and contextual features

I. INTRODUCTION

In recent years, most of the speech researchers are using unit selection procedures for speech synthesis. First the input text is normalized by expanding abbreviations, acronyms, numbers and all non-standard words. Recent findings suggest that the use of data-driven methods neural networks or statistical methods to to generate prosodic information and tom achieve naturalness and fluency in automatic speech synthesizers **[1].** Since duration is one of the important prosodic features, it is proposed to predict the duration.

Human brain **[2]** consist of three types of memories as long-term, short-term and mid-term which was studied in **[3, 4]** and **[5].** It has been shown that RNNs use short term memory. In RNN the connections between units form a directed cycle which allows it to exhibit dynamic temporal behavior. One or more feedback connections are used to pass output of a neuron in a certain layer to the previous layer(s). Due to the presence of cycles, it can not be divided into layers. RNN is more superior in learning many behaviors / sequence processing tasks / algorithms / programs compared to traditional machine learning methods.

A feed forward neural network is used to predict duration for Telugu **[6].** A Recurrent Neural Network (RNN) is used to predict prosodic information for Persian, Chinese and Mandarin **[7].** Recurrent data input also helps to smooth the output parameter tracks **[8].** RNNs inherently implement short-term memory by allowing the output of a neuron to influence its input either directly or indirectly via its effect on other neurons **[9].** It is obvious that cognitive processes and/or more practical applications will require higher-level architectures**.** This is a solid reason to investigate recurrent neural networks even if feed forward networks showed good results in many practical applications in different areas, from classification to time-series prediction. In the present work hence it is proposed to predict duration of syllable for Telugu with RNN approach since RNN is better in learning sequence processing tasks than simple feed forward neural network. Linguistic features are used as input nodes of RNN to learn duration rules of the syllable automatically and can be predicted duration of syllable at the output node.

II. RELATED WORK

Neural networks are very useful for applications like pattern recognition, data classification etc. through learning process. The RNN has been applied in a variety of areas including pattern recognition, classification, image processing, and combinatorial optimization and communication systems **[10].** A suitable algorithm should be considered for modeling the duration of basic units. RNNs have the ability of incorporating contextual or temporal dependencies in a natural way and also can include cyclic connections of the neurons. RNNs preserve some history of previous states through their recurrent links and accordingly they have been used widely in the processing of temporal patterns **[11].** Recurrent networks are built in such a way that the outputs of some neurons are fed back to the same neurons or to neurons in the preceeding layers **[12].** This helps in handling forward and backward coarticulation effect.

RNNs have an intrinsic dynamic memory and their outputs at a given instant reflect the current input as well as previous inputs and outputs which are gradually quenched. This has shown that how the synergistic combination of different local plasticity mechanisms can shape the global structure and dynamics of RNNs in meaningful and adaptive ways **[13].** Multilayer perceptron (MLP) and RNN are employed as local experts to discriminate time-invariant and time-variant phonemes, respectively **[14].** RNN can learn the temporal relationships of speech data and is capable of modeling time-dependent phonemes.

RNN can be trained to associate unknown input data to learned words **[15].** The neural network recognizer based on the static networks, such as MLP, and the dynamic networks like RNN **[16]** or Time Delay Neural Network (TDNN) **[17],** use parametric representation of the activation function. The exact label of the phoneme is determined at low level classification using RNN **[18].** MLP and RNN are employed as local experts to discriminate time-invariant and time-variant phonemes, respectively. RNN exhibits better performance in nonlinear channel equalization problem **[19].** To circumvent this difficulty, an adhoc solution has been suggested to back propagate the output error through this heterogeneous configuration.

RNN is a powerful connectionist model that can be applied to many challenging sequential problems, including problems that naturally arise in language and speech **[20].** However, RNNs are extremely hard to train on problems that have long-term dependencies, where it is necessary to remember events for many time steps before using them. However Temporal-Kernel Recurrent Neural Network (TKRNN) is very efficient for the long term dependencies. This is out of the scope of this work hence not discussed here.

It is observed that several types of features are used for duration modeling and more relevant work is briefly explained here. It employs a simple three layer RNN to learn the relationship between input prosodic features, with input syllable boundaries and output word-boundary information **[21].** Their experimental results show that the proposed Recurrent Fuzzy Neural Network (RFNN) can generate proper prosodic features including pitch means, pitch shapes, maximum energy levels, syllable duration and pause duration. The linguistic representation is usually a complex structure that includes information about the word sequence, Parts of Speech (POS) tags, prosodic phase information, fundamental frequency, energy and pause. The mixture of RNN expert’s type model provides robustness against changing the features in learning, but it lacks the ability to extract common patterns included in the sequences because of the independency of the local representation **[22].** The local representation is constructed into orthogonal units, while the global representation is also constructed into internal units using the connection weights between I/O units and internal units. Methods for processing speech data are described herein **[23].**

The input to the neural network consists of a set of features correspond to phonological, positional and contextual information which are extracted from the text **[24].** The relative importance of the positional and contextual features is examined separately. A two-stage duration model is proposed for improving the accuracy of duration. A multi-level prosodic model based on the estimation of prosodic features is considered **[25].** Different linguistic units to represent different scales of prosodic variations (local and global) at each level are used for syllable based duration modeling. Local and global variations are associated with phonological properties of these levels (coarticulation, syllabic structure, accentuation) and intermediate variations on a set of units larger than the syllable and + / - linguistically well defined (accentual group, interpausal group, prosodic group, intonational phrase, period, verbal construction, discourse sequence, ...) and associated with + / - linguistic factors : physiological (f0 declination), modalities (questions, ...), syntactical (prosodic contrasts related to some specific syntactical sequence), semantic (informational structure) and discourse. For duration, a phone-based **[26]** or a syllable-based **[27,28]** representation is considered.

A syllable based duration model based on multi-level context-dependent analysis is proposed **[30**]. In contrast to models based on modeling durational features on a single linguistic unit (phoneme, syllable), the proposed approach shows several advantages like distinguishing several linguistic units in the representation of durational features variation enables to explicit the superposition of prosodic forms jointly observed on a given unit, 2) each prosodic level (speech rate, duration syllabic residual, ...) can be modeled and controlled independently from each other and 3) estimate the set of linguistic features affecting each linguistic unit independently. In this experiment low-level linguistic features such as location features (position of a given unit within higher level units), weight features (number of observations of a given linguistic unit within higher level units) and phonological features (syllabic structure and prominence) are used.

A combination of the constraints and statistical analysis in the acquisition of the multiword acquisitions is outlined **[29].** Word forms in inflectional languages encode rich morpho-syntactic information constraining the possible syntactic structures. This type of information is useful in extraction of linguistic knowledge by means of M/C learning and statistical learning methods without resorting to parsing. Some approaches also make use of basic linguistic knowledge in the form of a heuristic method using language specific character frequencies plus language specific lists of function words and word endings **[30].** Common to all of these approaches is that the granularity of language identification is either a sentence or at most a word.

III. SPEECH DATABASE

In the current work Tv9 male speaker's speech is recorded. Since Telugu is syllabic in nature duration is predicted for syllable. Speech production as well as perception of Telugu can be considered as syllable like units. Also syllable like units capture some coarticulation effects. Syllable-like units considered are V, CV, CCV, CCVC and CVCC, where C is a consonant and V is a vowel. The database consists of Tv9 news data in Telugu language. Syllables of the form CV, CVC, CCVC, CVCC are extracted from text. The speech signal is sampled at 16 KHz sampling rate and encoded as 16-bit data. The speech utterances are manually transcribed into text using WX notation. Telugu has a character set of 56. These can be represented as V, CV and CCV forms. The TV9 speech is organized as syllables, words and phrases.

IV. DURATION MODELING

In continuous speech, different factors affect the duration of the basic units. They are classified into phonological, positional and contextual factors. Duration of syllable may be influenced by the category of the vowel present in the syllable, the category of the consonant(s) associated with the vowel and position of the vowel etc. Duration variations occur based on the positions of the basic units like word initial position, word final position, phrase boundary, sentence ending position etc. Similarly contextual variations occur due to the influence of the preceding and following units on the present unit.

RNN architecture **[24]** consists of three layers like input layer, hidden layer and output layer. At the input layer 25 input nodes are given. Output of input nodes is passed to hidden layer which consists of 40 hidden nodes and at the output layer 1 output node is connected. The activation function tan h is used at hidden layer. The most widely used training algorithm for RNN is the so called error back propagation. The aim of the algorithm is to adjust the weights from the output units to the hidden layer units and in turn from the units in the hidden layer to the input units to minimize the discrepancy between the network's output and its target, desired output. In back propagation this is done by propagating the error (i.e., the network's output for a given training vector (t-o) where t is a target vector and o is an output vector which is subtracted from the target, or vice versa back to the network in such a way that the weights are gradually adjusted to optimal values. In this work the objective is to adjust the weights of the network to minimize the mean squared error of each syllable's duration**.** This process is not deterministic and the networks do not always converge to the same solution.

Table-1: Details of features considered for RNN

Features at different levels	Details of Features	No. of nodes
Position of Syllable in the phrase	of the syllable from beginning of	1
	of the syllable from ending of	1
	Syllables in phrase	1

Position of Syllable in the word	n of the syllable from beginning of n of the syllable from ending of word Syllables in phrase	1 1 1
Syllable identity	Segments of syllable	4
Context of Syllable	us Syllable ving Syllable t Syllable	4 4 4
Syllable nucleus	1. Position of nucleus 2. No. of segments before nucleus 3. No. of segments after nucleus	1 1 1

In this work linguistic features like lexical (syllable identity, syllable nucleus), positional and contextual features are used. Syllable position in a phrase, word, syllable identity, context of a syllable and syllable nucleus are considered as input features for RNN **[12]**. Details of features at different levels considered for RNN is shown in Table 1.

The experiment is done in two phases as training and testing with two sets of data as train set and test set respectively. In the training phase initially the duration of the syllables are found manually. For each syllable the features extracted from text is given as input vectors. The corresponding syllable durations which are measured manually are given as output to the RNN models and these models are trained for 100 epochs. The training error is estimated for different combinations of input features and is shown in Figure 2, Figure 3 and Figure 4. In the test phase the predicted syllable duration is compared with the corresponding syllable from the test data. The difference between the actual duration and predicted duration is estimated as duration deviation. The deviation of duration for different syllable classes is shown in Table 2.

V. RESULTS

The output of RNN with different input features are shown below in Figure 1, Figure 2 and Figure 3.

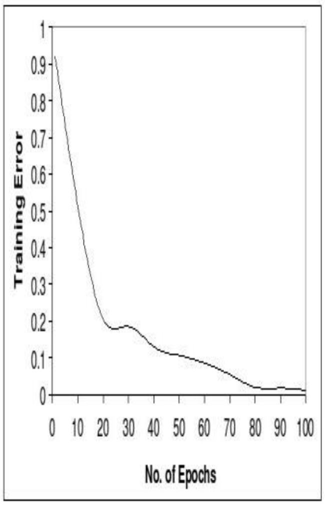

Figure 1: Trained on input features (Lexical + Positional + Contextual}

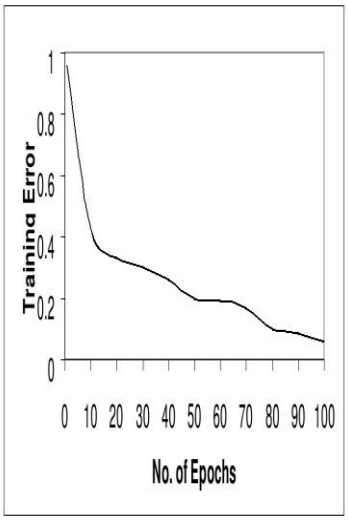

Figure 2: Trained on input features (Lexical + Contextual)

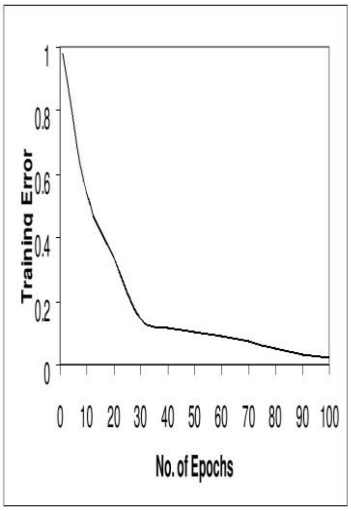

Figure 3:Trained on input features (Lexical + positional)

In Figure 1, is shown that the training error is decreasing to zero when RNN is trained with lexical, positional and contextual input features. In Figure 2, the training error is not reaching zero where RNN is trained with lexical and positional input features. It shows that these input features are not sufficient to predict duration correctly. In Figure 3, the training error is reaching approximately nearer to zero with lexical and contextual input parameters. Also at the beginning the training error is decreasing slowly compared training error with lexical, positional and contextual features. From these experiments it is clear that the combination of lexical, positional and contextual input features are useful for better duration prediction.

Table 2. Syllable duration deviation for different classes of sounds

Type of Syllable class	Percentage of duration deviation from original to predicted
Fricatives	6
Stops	4
Affricates	8
Nasals	2
Liquids	7
Labials	4
Alveolars	6
Palatals	4
Dentals	3
Velars	7
Voiced	2
Unvoiced	8

V. CONCLUSIONS AND FUTURE WORK

A Recurrent Neural Network is used for predicting the syllable duration. Duration values are predicted based on Phonological, positional and contextual information of syllables at phrase and word levels. It is observed that the duration values predicted are similar to values predicted by **[2,3].** The performance of neural net is evaluated with error values by using different combinations of input features. These error values are predicted by using the difference between actual duration and predicted duration. The values are shown in Table 2 and it is observed that the deviation of duration is in the decreasing order of voiced sounds, nasals < dentals < palatals, labials, stops < fricatives, alveolar < velars, liquids < affricates and unvoiced sounds. In future, performance can be improved by considering accent and prominence of syllable as additional feature vectors. Also it is proposed to study by considering some more additional features as input features. In RNN instead of taking all input features at single layer as input, a hierarchical approach with increasing size of the basic unit is to be tried.

REFERENCES

[1] S.Chen, S.Hwang and Y.Wang, An RNN based prosodic information synthesizer for Mandarin Text to Speech, Proc. Of ICASSP, Vol.6, Issue 3, PP:226-239, May 1998.

[2] Md. Khalilur Rhaman, Recurrent Neural Network Classifier for Three Layer Conceptual Network and Performance Evaluation, Proceedings of 11th International Conference on Computer and Information Technology (ICCIT 2008) 25-27

December, 2008, Khulna, Bangladesh, PP: 747- 752, 2008. Journal of Computers, Vol.5, No.1, PP:40-48, January 2010.

[3] E. Tulving and F. I. M. Craik (Editors), "The oxford handbook of memory."University press, New York, PP: 120-121, 1990.

[4] T. K. Landauer, "How much do people remember? Some estimates of the quantity of learned information in long term Memory." *Cognitive Science*, Vol.10, Issue 4, PP.477–493, 1986.

[5] R.L. Buckner, "Beyond HERA: Contributions of specific prefrontal brain areas to long-term memory retrieval, *Psychon Bulletin Review*, Vol.3, PP.149–158, 1996.

[6] K.Sreenivasa Rao and B. Yegnanarayana, Modeling syllable duration in Indian languages using neural networks, Proc. of IEEE ICASSP, Quebec, Canada, 17-21 May 2004, Vol:5,PP: 313-316, 2004.

[7] Farrokhi, Ali / Ghaemmaghami, Shahrokh / Sheikhan, Mansur, Estimation of prosodic information for Persian text-to- speech system using a recurrent neural network, Speech Prosody , Nara, Japan, March 23-26, PP: 475-478, March 2004.

[8] M.Riedi, A neural network based model of segmental duration for speech synthesis, Proc. of Eurospeech,PP:599-602, 1995.

[9] Utku Salihoglu, Toward a Brain-like Memory with Recurrent Neural Networks,Ph.D thesis, 2009 .

[10] Stelios Timotheou, The Random Neural Network: A Survey, The Computer Journal, Vol:53, Issue 3, PP: 251-267, 2010.

[11] B.Yegnanarayana, Artificial Neural Networks. New Delhi, India: Printice-Hall, 1999.

[12] W. N. Campbell, Analog I/O nets for syllable timing, Speech Communication, vol. 9, pp. 57–61, Feb. 1990.

[13] Andreea Lazar1, Gordon Pipa1,2 and Jochen Triesch, SORN:a self-organizing recurrent neural network, Computational Neurascience, Edited by: HavaT.Siegelmann, University of Massachusetts Amherst, USA, Front.Comput. Neurosci.3:23, 2009. .

[14] N.Uma Maheswari, A.P.Kabilan , R.Venkatesh, Speech ecognition System Based On Phonemes Using Neural Networks, IJCSNS International
Journal of Computer Science and Network Security, Vol.9, No.7, July 2009.

[15] Amrouche a; A. Taleb-Ahmed b; J. M. Rouvaen c; M. C. E. Yagoub, Improvement of the speech recognition in noisy environments using a nonparametric regression, International Journal of Parallel, Emergent and Distributed Systems, 24: 1, 49 - 67, 2009.

[16] Y.A. Alotaibi, Investigation of spoken Arabic digits in speech recognition setting, Informatics and Computer Science, 173 , PP. 105–139, 2005.

[17] A. Waibel, T. Harazawa, G. Hinton, K. Shakano, and K.G. Lang, Phoneme recognition using time delay neural networks, IEEE Trans.ASSP, Vol. 37, PP:328–339, (1989).

[18] N.Uma Maheswari, A.P.Kabilan, R.Venkatesh, Speaker Independent Phoneme Recognition Using Neural Networks, Journal of Theoretical and Applied Information Technology, Vol.6, No.2, PP:230- 235, 2009.

[19] Susmita Das, IEEE Member, A Novel Cascaded Nonlinear Equalizer Configuration on Recurrent Neural Network Framework for Communication Channel, Proceedings of the World Congress on Engineering 2009, Vol. I WCE 2009, July 1 - 3, 2009, London, U.K.

[20] Ilya Sutskever; Geoffrey Hinton, Neural networks : the official journal of the International Neural Network Society,Vol.: ISSN: 1879-2782 ISO Abbreviation: Neural Networks Publication, Nov. 2009.

[21] Chin-Teng Lin, *Senior Member, IEEE*, Rui-Cheng Wu, Jyh-Yeong Chang, and Sheng-Fu Liang, A Novel Prosodic Information Synthesizer Based on Recurrent Fuzzy Neural Network for the Chinese TTS System, IEEE Trans. On Systems, Man, and Cybernetics, Part B: Cybernetics, Vol.34, No.1, February 2004.

[22] Jun Namikawa and Jun Tani, Building Recurrent Neural Networks to Implement Multiple Attractor Dynamics Using the Gradient Descent Method, AdvancesinArtificialNeuralSystems, Vol. 2009, Article ID 846040,11 Pages.

[23] Ying, Zhiwei and Shi, Xiaohua, US Patent 7136802 - Method and apparatus for detecting prosodic Phrase break in a text to speech (TTS) system, US Patent Issued on November 14, 2006.

[24] K.Sreenivasa Rao and B.Yegnanarayana, Modeling durations of syllables using neural networks, Computer Speech and Language, Vol.21, Issue 2, Pages: 282-295, April 2007.

[25] Nicolas Obin, Xavier Rodet, Anne Lacheret-Dujour, A Multi-Level Context-Dependent Prosodic Model Applied to Durational Modeling, Interspeech 2009, Brighton, UK, pp.512-515, 2009.

[26] B. Gao, Y. Qian, Z. Wu, and F. Soong, "Duration refinement by jointly optimizing state and longer unit likelihood," Proc. of Interspeech, Brisbane, Australia, 2008.

[27] S.H. Chen, W.-H. Lai, and Y.-R. Wang, "A new duration modeling approach for mandarin speech, IEEE Trans. On Speech and Audio Processing, Vol. 11, no. 4, pp. 308–320, 2003.

[28 J. Latorre and M. Akamine, "Multilevel parametric-base f0 model for speech synthesis," Interspeech, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008, Brisbane, Australia, 2008.

[29] Maciej Piasecki and Adam Radziszewski, Aspects of natural language processing, Lecture notes on Computer science, Information Systems and Applications, incl. Internet/Web, and HCI , Vol. 5070, Marciniak, Malgorzata; Mykowiecka, Agnieszka (Eds.) , XII, 449 p., 2009, Springer Verlag.

[30] Harald Romsdorfer, Polyglot Text-to-Speech Synthesis Text Analysis & Prosody Control, Spech Communication, vol:49, PP:697-724, 2009.

BIOGRAPHY

P.N.Girija is presently working as Professor in the School of CIS, University of Hyderabad, Hyderabad. Her research areas are Speech Recognition, Speech Synthesis and Human Computer Interaction. she has published nearly eighty papers in various national and International journals and conferences. She has presented papers in several national and international conferences. She visited School of Computer Science, Carnegie Mellon University, Pittsburgh, U.S.A. as a visiting scholar during June-August 2004. She chaired several sessions like COCOSDA, NTU, Singapore etc. She completed sanctioned research projects from DST AICTE, UPE etc.

10.15680/ijircce.2015.0302017