Duration Modeling Using Multi-model Based on Positional Information

Vempada Ramu Reddy and Krothapalli Sreenivasa Rao

School of Information Technology,
Indian Institute of Technology Kharagpur
Kharagpur - 721302, West Bengal, India
ramu.csc@gmail.com, ksrao@iitkgp.ac.in

Abstract. This paper proposes prediction of syllable durations by developing multi-models using positional information. The proposed multi-model consists of four models used for predicting the durations of syllables. Among them, one of the models is used for predicting the durations of syllables present in mono-syllabic words, and the remaining three models are meant for predicting the durations of syllables present at initial, middle and final positions of polysyllabic words. In this study, (i) linguistic constraints represented by positional, contextual and phonological features and (ii) production constraints represented by articulatory features are used for predicting the duration patterns. Feed-forward Neural Networks (FFNN) are used for developing the duration models using above mentioned features. It was found, that the prediction accuracy is improved using multi-models compared to single duration model.

Keywords: Multi-models, Duration prediction, Prediction accuracy, Feed-forward neural networks, Linguistic and Production constraints.

1 Introduction

Duration plays a vital role in human speech communication. The sequence of syllable durations is defined as duration patterns. Variation in duration patterns provide naturalness to speech. Human hearing system is highly sensitive to the variations in duration patterns. Hence, while developing speech synthesis systems, acquisition and incorporation of the duration knowledge is very much essential [1].

In speech signal, the duration of each unit is dictated by the linguistic and production constraints of the unit [2] [3]. Duration models are developed using linguistic constraints represented by positional, contextual and phonological (PCP) features, and production constraints represented by articulatory (A) features. From here onwards the combination of features representing linguistic and production constraints is referred as PCPA features [1][4]. In this study, multi-model based approach is proposed for improving the prediction accuracy of durations of syllables. Multi-models are explored in this work based on

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 404–409, 2013.

positional aspects of syllables at the word level. The main reason for proposing multi-models for duration prediction is to avoid the bias towards mean during the prediction of syllable durations by single neural network. From the speech corpus, we have observed that durations of syllables has strong influence on their position in the word. Therefore, if we model the durations of syllables separately, based on their position in the word, the bias problem imposed by single neural network may be reduced. The implicit knowledge of duration is usually captured by using modeling techniques. Neural networks are used in this work to capture the underlying interactions that exist between input and output features [5] [3] [6].

The paper is organized as follows: The speech database used for modeling the syllable durations is presented in Section 2. Section 3 describes the features used for predicting the duration patterns. Performance of the proposed multi-model along with the single duration model using neural networks is explained in Section 4. Conclusions of this paper are presented in Section 5.

2 Speech Database

The text utterances of speech database used for this study are collected from Bengali Anandabazar newspaper, various text books and story books which covers wide range of domains. The collected text corpus covers 7762 declarative sentences with 4372 unique syllables and 22382 unique words. The text is recorded with a professional female artist in a noiseless chamber. The duration of total recorded speech is around 10 hrs. The speech signal was sampled at 16 kHz and represented as 16 bit numbers. The speech utterances are segmented and labeled into syllable-like units using ergodic hidden Markov models (EHMM). For every utterance a labeled file is maintained which consists of syllables of the utterance and their timing information. The percentage of different syllable structures present in the database are V(8.20%), VC(3.50%), VCC (0.20%), CV(50.41%), CVC(32.26%), CVCC(1.05%), CCV(2.50%), CCVC(1.77%) and CCCV(0.11%), where C is a consonant and V is a vowel. It is observed that durations of syllables in the database vary from 50 to 560 ms with mean and standard deviations 212.9 ms and 80.6 ms, respectively.

3 Features for Modeling the Durations

It is known that there exists some inherent relationship between linguistic and production constraints of speech to the duration variation patterns in speech [1] [7]. The linguistic and production constraints are represented by positional, contextual, phonological and articulatory (PCPA) features [8] [6] [9] . The positional features are further classified based on syllable position in a word and sentence, and word position in a sentence. The detailed list of linguistic and production constraints are given in Tables 1 and 2, respectively.

Table 1. List of factors affecting duration patterns of syllables, features representing the factors and number of nodes needed for neural network to represent the features

Factors	Features	# Nodes
Syllable position in the sentence	Position of syllable from beginning of the sentence Position of syllable from end of the sentence Number of syllables in the sentence	3
Syllable position in the word	Position of syllable from beginning of the word Position of syllable from end of the word Number of syllables in the word	3
Word position in the sentence	Position of word from beginning of the sentence Position of word from end of the sentence Number of words in the sentence	3
Syllable identity	Segments of the syllable (consonants and vowels)	4
Context of the syllable	Previous syllable Following syllable	4 4
Syllable nucleus	Number of segments before the nucleus Number of segments after the nucleus Number of segments in a syllable	3

Table 2. List of articulatory features

Features	Description
vlen	Length of the vowel in a syllable (short, long, dipthong and schwa).
vheight	Height of the vowel in a syllable (high, mid and low).
vfront	Frontness of the vowel in syllable (front, mid and back).
vrnd	Lip roundness (no rounding and rounding).
ctype	Type of consonant (stop, fricative, affricative, nasal, and liquid).
cplace	Place or position of the production of the consonant (labial, alveolar, palatal, labio-dental, dental and velar).
cvox	Whether consonant is voiced or unvoiced (voiced and unvoiced).
asp	Whether consonant is aspirated or unaspirated (aspirated and unaspirated).
nuk	Whether consonant with nukta or not nukta (withnukta and withoutnukta).
fph	Type of first phone in a syllable (vowel, voiced consonant, unvoiced consonant, nasal, semivowel, nukta and fricative).
lph	Type of last phone in a syllable (vowel, voiced consonant, unvoiced consonant, nasal, semivowel, nukta and fricative).

4 Proposed Multi-model Based Approach

In this study, multi-model is developed by dividing the syllables into four groups namely syllables in mono-syllabic words, syllables at initial, middle and final positions in polysyllabic words. Feed-forward neural networks (FFNN) are used in this work for developing multi-model. For each group of syllables, a separate model

is prepared using FFNN. Finally, the multi-model used in this work consists of 4 FFNNs developed by 4 groups of syllables mentioned above. After grouping the syllables, it is observed that syllables from mono-syllabic words represent only 5% of total syllables, and the remaining 95% is distributed as 35%, 25% and 35% syllables from initial, middle and final positions of the poly-syllabic words. The prediction performance of multi-model is compared with the single duration model developed using PCPA features. In this work, single duration model refers to single FFNN trained with PCPA features of all syllables together. The details of neural networks and the performance of multi-model are discussed in the following subsections.

4.1 Neural Networks

In this work, a four layer feed-forward neural networks (FFNN) [8] [10] with input layer, two hidden layers and output layer is used to model the durations of syllables. The structure of the four layer FFNN is shown in Figure 1. In this work, the data consists of 177820 syllables is used for modeling the durations of sequence of syllables. Different structures were explored to obtain the optimal four layer FFNN, by incrementally varying the hidden layer neurons in between 5 to 100. The details of neural network is given in Table 3.

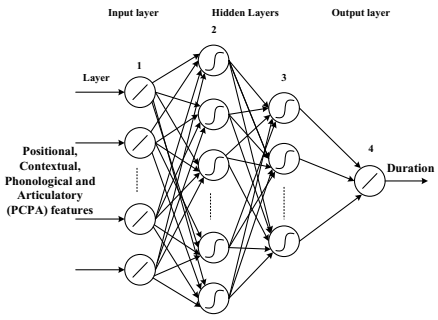

Fig. 1. Architecture of four layer feed-forward neural network

Table 3. Details of neural network

Non-linear activation function	tanh(s) where 's' is activation value
Training algorithm	Levenberg-Marquardt backpropagation
Train data	70%
Validation data	15%
Test data	15%

4.2 Evaluation

The prediction accuracy of the models is evaluated by means of objective measures such as average prediction error (μ), standard deviation (σ) and linear correlation coefficient ($\gamma_{X,Y}$). The computation of objective measures is given below:

$$D_i = \frac{|x_i - y_i|}{x_i} \times 100, \mu = \frac{\sum_i |x_i - y_i|}{N}, \sigma = \sqrt{\frac{\sum_i d_i^2}{N}}, \text{ and } \gamma_{X,Y} = \frac{V_{X,Y}}{\sigma_X . \sigma_Y}$$

$$\text{where } d_i = e_i - \mu, e_i = x_i - y_i, \text{ and } V_{X,Y} = \frac{\sum_i |x_i - \bar{x}| . |y_i - \bar{y}|}{N}$$

where x_i, y_i are the actual and predicted duration values, respectively, and e_i is the error between the actual and predicted duration values. The deviation in error is d_i, and N is the number of observed duration values of the syllables. σ_X, σ_Y are the standard deviations for the actual and predicted duration values respectively, and $V_{X,Y}$ is the correlation between the actual and predicted duration values.

The prediction performance of individual models of multi-model ($1-4$ rows), overall multi-model performance (5^{th} row) and performance of single model (6^{th} row) is given in Table 4. Column 1 of Table 4 shows type of model used for prediction. Columns 2-6 indicates the percentage of syllables predicted within different deviations from their actual duration values and columns 7-9 indicates the objective measures.

Table 4. Performance of FFNN based multi-model for predicting the duration values of syllables

Models	**% Predicted syllables within deviation**					**Objective measures**		
	2%	5%	10%	15%	25%	μ (ms)	σ (ms)	γ
Mono	8.33	19.02	38.36	56.71	77.22	55.46	44.18	0.83
Initial	10.39	23.85	42.07	62.20	81.62	26.59	22.22	0.89
Middle	7.49	18.31	31.25	50.16	73.66	29.60	24.74	0.83
Final	8.51	19.98	37.61	54.58	76.57	34.93	26.40	0.84
Multi	**8.90**	**20.86**	**37.60**	**56.23**	**77.63**	**31.71**	**25.42**	**0.86**
Single	7.96	18.88	35.14	50.63	72.56	39.04	35.09	0.83

From Table 4, it is observed that the performance of the model developed by using syllables representing initial position in word performed better than other models. From this we can hypothesize that initial position syllables are more discriminating than other syllables. It is also observed that the average prediction error of mono-syllables is high compared to others. The high average error of mono-syllables is mainly due to insufficient amount of syllables present in the database for training. However the overall performance of multi-model is outperformed compared to single duration model developed by using all syllables. From this hypothesis, we can conclude that the prediction accuracy of durations is improved by dividing the syllables and developing multi-model based

on the syllable position in the word. It was observed in the database that the mean durations of mono-syllables, initial, middle and final position of syllables in words vary greatly. Therefore, separating the syllables based on their position and developing the multi-model eliminated the biasses of one group of syllables towards mean values of other groups. This resulted in the improvement in the performance compared to single model.

5 Conclusions

In this work, prediction of durations of syllables is carried out using multi-model based approach. Multi-model based on syllable position in a word is developed by using neural networks. Linguistic and production constraints represented by positional, contextual, phonological and articulatory features are used for modeling the duration patterns. It is observed, that the performance of proposed multi-model based approach is performed better compared to single model. In future, prediction of durations can be analyzed by developing multi-models based on production aspects of speech segments.

References

1. Reddy, V.R., Rao, K.S.: Better human computer interaction by enhancing the quality of text-to-speech synthesis. In: Proc. Int. Conf. Intelligent Human Computer Interaction (IHCI), IIT Kharagpur, India, pp. 1–6 (December 2012)
2. Rao, K.S., Yegnanarayana, B.: Modeling durations of syllables using neural networks. Computer Speech and Language 21, 282–295 (2007)
3. Sreenivasa Rao, K., Mahadeva Prasanna, S.R., Yegnanarayana, B.: Two-stage duration model for Indian languages using neural networks. In: Pal, N.R., Kasabov, N., Mudi, R.K., Pal, S., Parui, S.K. (eds.) ICONIP 2004. LNCS, vol. 3316, pp. 1179–1185. Springer, Heidelberg (2004)
4. Reddy, V.R., Rao, K.S.: Intonation Modeling Using Linguistic, Production and Prosodic Constraints for Syllable based TTS Systems. Procedia Engineering, Elsevier 38, 2772–2783 (2012)
5. Yegnanarayana, B.: Artificial Neural Networks. Prentice-Hall, New Delhi (1999)
6. Reddy, V.R., Rao, K.S.: Intonation Modeling using FFNN for Syllable based Bengali Text To Speech Synthesis. In: Proc. Int. Conf. Computer and Communication Technology, MNNIT, Allahabad, pp. 334–339 (2011)
7. Rao, K.S., Yegnanarayana, B.: Intonation modeling for Indian languages. Computer Speech and Language 23, 240–256 (2009)
8. Reddy, V.R., Rao, K.S.: Two-Stage Intonation Modeling Using Feedforward Neural Networks for Syllable based Text-to-Speech Synthesis. Computer Speech and Language 27, 1105–1126 (2013)
9. Ramu Reddy, V., Sreenivasa Rao, K.: Intensity Modeling for Syllable Based Text-to-Speech Synthesis. In: Parashar, M., Kaushik, D., Rana, O.F., Samtaney, R., Yang, Y., Zomaya, A. (eds.) IC3 2012. CCIS, vol. 306, pp. 106–117. Springer, Heidelberg (2012)
10. Tamura, S., Tateishi, M.: Capabilities of a Four-Layered Feedforward Neural Network: Four Layers Versus Three. 8, 251–255 (1997)