Phonetic words duration simulation using Deep Neural Networks

Alexander Shipilo
Saint-Petersburg State University, Russia
https://doi.org/10.36505/ExLing-2016/07/0037/000296

Abstract
Deep Neural Networks (DNN) are widely used in speech prediction and speech modeling. The current paper describes the implementation of DNN for the task of duration prediction of speech units (allophones and syllables that form the structure of phonetic word, intonation phrase). It is well-known that numerous factors influence the duration of segments. However, the level of confidence of characteristics differs significantly. It was found that deep neural network that predicts allophones duration shows better results than the network that predicts the duration of syllables.
Key words: deep NN, duration modeling, phonetic words.

Introduction
One of the challenging tasks in text-to-speech systems is the problem of duration modeling of speech units. Despite recent research refers to the problem of lengthening and shortening the speech units, unit selec-tion systems demonstrate better naturalness (Lobanov, Tsirulnik, 2007, 2008).

The duration of speech segments varies significantly depending on the position within intonation unit, phonetic word, the number of elements in the speech unit (Svetozarova, 2014). Each allophone unit has its own intrinsic duration value. It is known that a lot of factors influence the segment duration.

Python Toolkit for Deep Learning (PDNN) was used in the current research. The general architecture of the developed system is shown on the fig. 1.

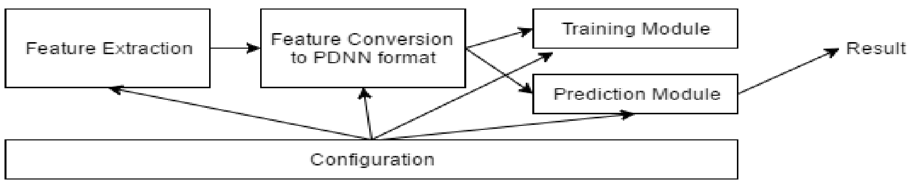

Figure 1. The architecture of the system.

Material

The Corpus of Professionally Read Speech (CORPRES) was used in the current research (Skrelin et al., 2010). During the pilot experiment the recordings of one female speaker (approx. 6 hours of speech, 155591 allophones, 61591 syllables) were chosen. Each recording has following manual checked annotation level:

F0 marks (stylized according to (Skrelin, Kocharov, 2009)
Ideal transcription
Real transcription
Word boundaries
Pitch movements
Boundaries of intonation units

CORPRES doesn't contain syllable and phonetic words annotations. To generate these levels a python script was created. Syllable boundaries were estimated according to Shcherba syllabification theory (Matusevich, 1976). According to allophone boundaries level a new level was automatically generated. It contains the allophone boundaries and boundaries of non-phonemic units (vowel insertions etc.).

The segment duration in the material is pre-processed. Each allophone segment is normalized according to the tempo coefficient estimated by

$$T = (D1 / N) / (D2 / N),$$

where T – tempo coefficient, D1 – sum of average duration values of allophones within intonation unit, D2 – sum of real durations within intonation unit.

Experiments

Four experiments were performed. The first two deal with syllable duration prediction (models 1, 2), others – with the allophones duration prediction (models 3,4). Let us consider the experiment techniques.

Unfortunately, the prediction of real duration of a segment is a rather difficult task. To simplify it, rounded values were predicted. Model 1 recognizes the percent deviation from the average of all syllables in the material, model 3 – the percent deviation of the required allophone. For example, let us consider the segment that is lengthened by 10 percent. This value was rounded to the nearest possible percent deviation value. If the required value is 110 %, the required coefficient is rounded to the nearest possible value accurate to 25 percent (e.g. 110 to 100 %, 120 to 125 % etc.). Table 1 shows the features that were used in the model.

Models 3,4 predict the rounded number that is required to multiply by the minimum level of auditory perception that is equal to 30 ms.

Each model consists of two hidden layers, each layer contains 2048 elements.

Table 1. The features for the models.

Features	
Model 1,2	Model 3,4
	Allophone (previous, current, following)
	Number of allophones in the syllable
Syllable index from the beginning/end of phonetic word	Syllable index from the beginning/end of phonetic word
Number of phonetic words in intonation unit	Number of phonetic words in intonation unit
Number of syllables in intonation unit	Number of syllables in intonation unit
	Allophone index from the beginning/end of syllable
Phonetic word index from the beginning/end of intonation unit	Phonetic word index from the beginning/end of intonation unit
The reduction level of a syllable	The reduction level of a syllable
The pitch movements within syllable, phonetic word, intonation unit	The pitch movements within allophone, syllable, phonetic word, intonation unit

Results and discussion

Table 2 demonstrates the results of the experiments.

Table 2. Results.

Model	Prediction accuracy, %
1	20
2	51
3	45
4	79

As we can see from the table 1, model 4 shows the best result, model 1 – the worst. Models that simulate syllable durations show worse results,

than models that simulate the allophone ones. Models 2 and 4 show better results (51 and 79 percents) in comparison to models 1 and 3. The reasons for it are the fact that (1) the deviation depends on the duration of the average of the target element, (2) the deviation is the relative characteristic. Let us consider the average unstressed vowel allophone (for example /u/) equals 50 ms. In this case ten percent lengthening means that the duration changes by 5 ms. On the other hand, the ten percent change of the stressed allophone of phoneme /a/ (the average duration in the material is 109 ms) means that the duration changes by approximately 11 milliseconds. If we predict the real allophone duration (models 2, 4), the problem of differences in averages disappears.
The results confirm the hypothesis that the selected features can be used as predictors of segment durations, but neural net provides no information about the rate of confidence of the features. To answer this question additional study is required.

References

Lobanov, B.M, Tsirulnik, L.I., Rules of Speech Corpus Segmentation into Phonetic Units and the Strategy of Unit Selection in Speech Synthesis, http://www.dialog-21.ru/digests/dialog2007/materials/html/60.htm

Lobanov, B.M, Tsirulnik, L.I., Computer Synthesis and Speech Cloning, Minsk, 2008 / in RussianMatusevich, M.I. 1976. Modern Russian Language. Phonetics (in Russian) Sovremennij Russkij Yazik. Phonetika

Matusevich, M.I., Modern Russian Language. Phonetics, 1976 / in Russian/ Sovremennij Russkij Yazik. Phonetika

Skrelin P., Kocharov D., Automatic processing of prosodic design of the utterance: relevant prosodic features for automatic interpretation of intonation model, 2009, AP-2009, Saint-Petersburg / in Russian.

Skrelin, P., Volskaya, N., Kocharov, D., Glotova, O., Evdokimova V. CORPRES - Corpus of Russian professionally read speech. In: Sojka, P., Horák A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp., 392-399. Springer, Heidelberg (2010)

Svetozarova N.D., “Short” stressed vowels in the Russian language, Issues in Phonetics 6, 2014 / in Russian.