EMPHATIC SPEECH PROSODY PREDICTION WITH DEEP LSTM NETWORKS

Slava Shechtman, Moran Mordechay

IBM Research AI, Haifa Research Lab, Haifa, Israel

ABSTRACT

Controllable generation of emphasis in speech is desirable for expressive TTS systems utilized in various dialog applications. Usually such models remain voice-specific and the strength of emphasis can't be readily controlled. In this work we present a flexible emphatic prosody generation model based on Deep Recurrent Neural Networks (DRNN) for controllable word-level emphasis realization. The word emphasis DRNN model was trained on syllable-level piece-wise linear prosodic trajectory parameters. A special data preprocessing technique was introduced to enable emphasis strength control, allowing to generate emphatic prosody trajectories of various strength. Additionally, we trained a DRNN model generating a sentence-level emphasis, i.e. producing whole sentences in forceful, decisive manner. Both models preserve quality and naturalness of the baseline TTS output.

Index Terms— TTS, speech synthesis, expressive speech synthesis, emphasis, emphatic speech, Deep learning, LSTM

1. INTRODUCTION

Much of information, carried by human speech, is beyond verbal and is conveyed by speech prosody. Besides carrying a certain prosodic pattern characteristic of a given language, the prosody determines emotional state and attitude of speakers and also helps to bring clear messages to listeners by distinguishing more important speech portions from the rest. The latter is usually realized by means of word emphasis. The word emphasis is either applied deliberately to convey certain speaking style or used pragmatically to focus attention on particular words or the ideas associated with them. Doing so can change or clarify the meaning of a sentence. Clearly, the controllable generation of word emphasis in speech is useful for high quality Text To Speech (TTS) systems utilized in various dialog applications, e.g. virtual personal or sales assistants.

Emphasized words usually manifest themselves by various prosodic and acoustic features, such as pauses before and/or after the word, a slower speaking rate in the word, a higher energy in the word, an increased activity in the intonation, or a combination of these features. Emphasis production is language and speaker dependent. It is also dependent on long semantic contexts, and even in the same context the same speaker can realize it with various extent.

Enabling word emphasis in TTS frequently requires dedicated audio data collection [1], which is constrained by a limited amount of emphasized words that can be uttered naturally in a single sentence. An alternative data collection approach is usually based on a manual effort to annotate emphasized words within an existing voice corpora [2], but the strength of such emphasis realizations might be modest compared to the dedicated emphasis datasets. The latter approach would benefit a lot from a robust and controllable way to modify (usually, increase) the emphasis strength if required by an application.

The word emphasis modeling has been extensively explored for unit-selection systems [1][3], in which the emphasis attributes mostly influenced the selection of concatenated units, and for HMM-based parametric synthesis [2][4][5]. Specifically, missing data [4] and continuous emphasis control [5] challenges were tackled in HMM-based TTS systems by modification of the decision tree based clustering procedures.

To our best knowledge, the emphasis has not yet been explicitly explored using recent state-of-the-art technologies for prosody modeling (i.e. Deep Neural Networks, or DNN), and this is the topic of our work. The main goal of this research is the incorporation of a controllable word emphasis into high quality prosodic trajectories, directly applicable for state-of-the-art TTS systems, such as large-scale unit selection [4], high quality parametric DNN-based [6] or non-parametric DNN-based [7] synthesis. In addition, we want to explore a *sentence emphasis,* i.e. emphasizing a sentence as a whole, to make it sound in forceful, decisive manner. This emphatic speaking style might be usable in generic TTS to generate short key sentences.

The paper is structured as follows. First, we describe the underlying speech synthesis engine used in this work. Second, we elaborate on our controllable word emphasis model. Then, we present a whole-sentence emphasis mode. Finally, experimental results will be presented.

2. SPEECH SYNTHESIS ENGINE

The underlying speech synthesis engine used in this work is the IBM concatenative unit-selection system with its prosody predicted with a Bidirectional Recurrent Neural Network with Long Short-Term Memory Units (BiRNN-LSTM) [8]. The predicted prosody target served for unit

selection. In addition, post-selection signal modification by PSOLA [8] was performed to better fit the pitch targets at the emphatic areas and the duration targets in sonorant speech areas. The baseline network contains 3 bidirectional hidden layers (*65*, *55*, *45*), and the 4th fully connected linear layer that generates 4 outputs per TTS unit, i.e. the target duration, start pitch, end pitch and energy. The TTS units correspond to roughly 1/3 of a phone and result from forced-alignments with 3-state hidden Markov models.

The DRNN input features are comprised of *1-hot* coded categorical features (e.g., syllable stress, part of speech (POS), phrase type, etc.) and standard positional features (e.g., number of phones/words to/from a phrase/utterance boundary, etc.). Feature values are propagated down to the constituent units.

One of the meaningful input categorical features is a rule-based word prominence, which may take one of seven levels. The word prominence determination rules are based on such features as POS and a word position in the phrase. There are also specific rules for certain words and word sequences. Usually, function words receive low prominence values, while content words receive the two highest values, with the highest value mostly assigned to phrase final content words. However, there are some exceptions. Meaningful function words such as "every", "most" and "not" receive high prominence values.

3. WORD EMPHASIS PROSODY PREDICTION

The direct approach for word emphasis modeling is to extend the input feature-set (to the generic BiRNN-LSTM [8]) by a binary indicator feature. This approach would serve as a trivial reference to compare with (see below).

3.1. Prosody Trajectory Parameterization

To tackle the lack-of-data problem it is desirable to reduce the sequence resolution in the sequence-to-sequence emphatic prosody prediction task. Given a high quality generic prosody trajectory prediction of high resolution [8], we propose to decompose it, syllable-by-syllable, to a simple piecewise-linear trajectory [9] and a corresponding residual. Moving to the syllabic resolution, we remain with roughly 10-times shorter sequences that supposedly contain all the necessary information for emphasis (or stress) modeling, which is known to be syllable-based [9].

The proposed syllabic parameterization is applied just on a syllable nucleus, defined here as a sonorant portion of a syllable, surrounding its vowel, that contains also glides, liquids and nasals, e.g. *'r','w','l',m','n'*, but not fricatives, like *'v','z','g'*.

Let $p(t)$ be a continuous fine-grained piecewise-linear log-pitch trajectory of a syllable nucleus, connecting a sequence of N break points $\{(t_1, p_1), \dots, (t_N, p_N)\}$, constructed from the corresponding prosodic targets, evaluated per speech unit. The unit prosody targets are predicted with BiRNN-LSTM [8]. The time scale is normalized, i.e. t_n is in the range of [0,1], and the nucleus duration d is stored separately.

In our proposed parameterization, the log-pitch trajectory $p(t)$ is approximated as $\hat{p}(t)$, a piecewise linear curve with a single break point. The break point is selected to be the most prominent point on the log-pitch trajectory of the vowel (i.e. a point which is both a local and a global extremum, but not on the vowel boundaries). If the prominent point does not exist (e.g. the pitch trajectory is monotonous within the syllabic vowel) the breakpoint is selected to be the vowel mid-point. The normalized placement of the mid-point, t_{mid} is stored for the sake of the trajectory reconstruction.

Once the break point (t_{mid}, p_{mid}) is determined, the left and the right log-pitch linear approximations are evaluated by a linear regression of the uniformly sampled upper and lower parts of the log-pitch trajectory:

$$\begin{aligned} \boldsymbol{P}_{low} &= \{p(\tau k)\}_{k=0}^{\lfloor t_{mid}/\tau \rfloor}, \boldsymbol{T}_{low} = \{\tau k\}_{k=0}^{\lfloor t_{mid}/\tau \rfloor} \\ \boldsymbol{P}_{up} &= \{p(\tau k)\}_{k=\lfloor t_{mid}/\tau \rfloor}^{\lfloor 1/\tau \rfloor}, \boldsymbol{T}_{up} = \{\tau k\}_{k=\lfloor t_{mid}/\tau \rfloor}^{\lfloor 1/\tau \rfloor} \end{aligned} \quad (1)$$

For the sake of training, the stylized log-pitch trajectory is described by log-pitch differences at nucleus boundaries, with respect to p_{mid}:

$$\begin{aligned} \Delta p_{start} &= \frac{(\boldsymbol{P}_{low} - p_{mid})^T (\boldsymbol{T}_{low} - t_{mid})}{\|\boldsymbol{T}_{low} - t_{mid}\|^2} (-t_{mid}) \\ \Delta p_{end} &= \frac{(\boldsymbol{P}_{up} - p_{mid})^T (\boldsymbol{T}_{up} - t_{mid})}{\|\boldsymbol{T}_{up} - t_{mid}\|^2} (1 - t_{mid}) \end{aligned} \quad (2)$$

Once the approximated trajectory is obtained, the residual trajectory is evaluated as

$$\boldsymbol{r} = \{p(t_n) - \hat{p}(t_n)\}_{n=1}^{N}, \boldsymbol{t} = \{t_n\}_{n=1}^{N} \quad (3)$$

The prosodic parameters used for the emphasis model training comprise of four components: the nucleus log duration, $\log(d)$, the mid log-pitch, p_{mid} and the log-pitch boundary differences Δp_{start} and Δp_{end}. Other parameters, including residual log-pitch trajectory and mid-point placement are not modeled, but preserved for the reconstruction.

3.2. Controllable Data Preprocessing

The ultimate goal of the emphatic model is to learn a difference between the emphatic prosody realization and its corresponding neutral realization. We found experimentally that learning the direct difference [10] between the predicted neutral prosody [8] and the emphatic prosody, didn't help much, probably because of the adverse missing-data conditions.

A special target data preprocessing technique, described below, helped to attain an effective emphasis model.

Let $\{s(m)\}_{m=1}^{M}$ be an observed trajectory of a certain component of the syllabic target vector over time and let $\{s_{neu}(m)\}_{m=1}^{M}$ be a corresponding predicted neutral component, obtained from the baseline fine-grained prosody prediction [8]. (Here m is a running index of syllables in an utterance). We assign L_n to be a subset of indices in the

vicinity of the *n*-th syllable, sharing similar to the *n*-th syllable functionality (e.g. indices to syllables within the same prosodic phrase that have the same lexicographic stress). Then, we define a parametric family of reference α-trajectories $\left\{\left\{s_{ref}^{\alpha}(m)\right\}_{m=1}^{M}\right\}^{\alpha}$, for $-1 \le \alpha \le 1$:

$$s_{ref}^{\alpha}(m) = \begin{cases} \alpha \underset{q\in L_m}{med}\left(s_{neu}(q)\right) + (1-\alpha) \underset{q\in L_m}{\max}\left(s_{neu}(q)\right), \alpha \ge 0 \\ |\alpha| \underset{q\in L_m}{med}\left(s_{neu}(q)\right) + (1-|\alpha|) \underset{q\in L_m}{\min}\left(s_{neu}(q)\right), \alpha < 0 \end{cases} \quad (4)$$

where *med()* is either median or mean operator (see Table 1)

The target component preprocessing is performed by subtracting the reference α-trajectory with some experimentally determined parameter α. After the prediction, an α-trajectory with different (usually larger) α can be added back. The difference in pre- and post-processing allowed us to control word emphasis extent.

In our experiments the described data pre-processing/post-processing was applied on the log-duration and the mid log-pitch components of the syllabic target vector, as detailed in Table 1.

Table 1. **Data pre- and post-processing for normal emphasis (WE1) and strong emphasis (WE2) models**

	L_m	*med()*	*WE1, α*		*WE2, α*	
			pre-	post-	pre-	post-
log(*d*)	Same prosodic phrase, same binary stress	Median	0.0	0.1	0.0	0.1
p_{mid}	Same prosodic phrase, any stress	Mean, weighted by duration (d)	0.0	0.4	-0.9	-0.1

3.3. DRNN modeling

In this work we would like to remain within the same sequence-to-sequence BiRNN-LSTM prosody prediction framework, that resulted in the state-of-the-art generic prosody prediction [8]. Fortunately, the dimensionality of the baseline categorical and positional feature set of 336 features per sub-phonemic unit [8] can be reduced by removing features with the resolution higher then syllable and lower then prosodic phrase. The reduced set of 120 features is fed into a BiRNN-LSTM network (trained with the pre-processed syllabic targets), comprised of 3 stacked bi-directional LSTM layers (*20, 17, 14*).

3.3.1. Training

Training deep networks for word emphasis prosody prediction is harder than general prosody prediction task due to a small proportion of emphasized data within natural speech. To diminish the missing data problem, we performed two stage training. First, we trained the model from the scratch with respect to the weighted square-loss function with class-dependent weights, defined as follows: for stressed syllables in emphasized words the weight was *3*, for stressed syllables in other words the weight was *1,* and for unstressed syllables the stress was 0 (no contribution to training). Then we retrained the neural net with non-zero loss on stressed syllables within emphasized words only, using the above model as initial condition. We deployed a variation of stochastic gradient descent with early stopping for training at each stage.

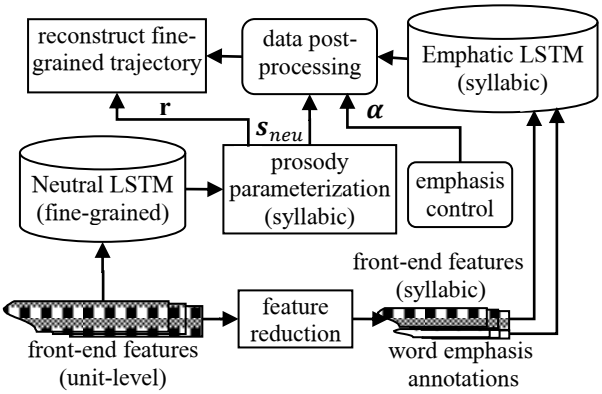

Figure 1. Word emphasis prediction

3.3.2. Prediction

The structure of hierarchical emphatic prosody generation is displayed on Figure 1. We first predict a fine-grained neutral prosody trajectory from the text-based front-end features.

The prediction of the 4-dimentional syllabic targets (*log-duration, mid log-pitch, delta-log-pitch* to *start* and *delta log-pitch* to *end*) is performed only for stressed (i.e. having either primary or secondary stress) syllables of the words, annotated as emphasized. Then we apply the post-processing and reconstruct the fine-grained trajectory by adding residual syllabic parameters, extracted from the neutral fine-grained trajectory.

4. SENTENCE-LEVEL EMPHASIS PREDICTION

In various scenarios for TTS one can identify key sentences (or *emphatic sentences*) that should be uttered in a forceful, decisive manner, e.g. virtual sales agent, auto-summarization, e-learning, etc.

If word emphasis annotation is not provided to TTS during the synthesis, we cannot directly apply the proposed word emphasis model for the emphatic sentences. As a simple workaround, one might try to apply the proposed word level emphasis with reduced strength to all the meaningful words in the sentence. Apparently, this naive approach did not work (see Section 5), so we had to learn the emphatic sentence prosody directly from the voice corpus, However, only 3% of our expressive voice corpus was annotated as emphatic sentences, so we had to deal with missing-data issue.

To indicate whether the current sentence is emphatic, a binary indicator feature was added to the DRNN input feature set. In addition, a rule-based word prominence feature (see section 2), was extended with one more category indicating the annotated emphasized words in the corpus during the training. During the synthesis, the rule-based prominence of all the meaningful words (i.e. having the two highest rule-based prominence levels) within the emphatic sentence are substituted by this new category. This

way we implicitly learned how to emphasize all the meaningful words in the emphatic sentences.

An example of meaningful words in an emphatic sentence is presented below in bold:

*It is **principally** the **viewing rates** which **decide** upon the **program** in the **private radio** and **television business**.*

To cope with the missing data problem, we augmented the training data by 5-fold duplication of the key sentences (with their annotations) in the corpus. The training was performed with 90% of data, including about 90% of original emphatic sentences with their duplications. The rest served for early-stopping during the training.

5. EXPERIMENTAL RESULTS

A concatenative speech database, consisting of approximately 20 hours of professionally recorded speech from a native female speaker of US English, has been used as data to train prosodic models to be deployed in the IBM unit-selection TTS [8]. Most of the corpus was comprised of fragments from audience addressed speeches, in which the speaker was instructed to read in a persuasive and lively manner. Based on the recorded speech, the corpus was annotated by 4 professional labelers with *emphatic sentence* and *emphasized word* labels (which sometimes overlap) *Emphasized word* labels that resulted from agreement of 1 out of 4 labelers were used for training the emphasis models. Since *emphatic sentence* labeling seems to be a more complicated task, agreement of 3 labelers was used in order to ensure high quality labels. There were approximately 26,000 emphasized words labeled (about 16% of the corpus) and approximately 600 emphatic sentences labeled (about 3% of the corpus), prior to the emphatic sentence duplication.

To evaluate the proposed systems, several subjective listening evaluations were conducted in a form of Mean Opinion Score (MOS) tests [11] with 40 out-of-corpus stimuli per system and 25-60 votes per stimulus, provided by 40-80 paid anonymous native speakers. Around 10% of subjects were removed as a result of the outlier rejection [11]. In addition to the neutral prosody reference model, (*Ref0)*, a default BiRNN-LSTM [8] with an extra binary word emphasis feature was trained (*Ref1*) as a reference for the proposed word emphasis prediction systems with various strength (*WE1*, *WE2*, see Table 1). In addition to the standard quality and naturalness MOS test [11], users were asked to assess the emphasis in annotated words (the text with emphasized words annotated, single word per sentence, was given). 1-5 scale was utilized for this test with some of values explained (1: neutrally spoken, 3: somewhat emphasized, 5: definitely emphasized) The evaluation scores are reported along with their 95% confidence interval in Table 2. The bold results are statistically significant ($p<0.05$) compared to the reference system (Ref0). One can observe that for WE1 both the emphasis effectiveness and the quality improved with statistical significance. For WE2 the emphasis effectiveness improved even more, while the quality degraded non-significantly compared to Ref0. Additionally, the proposed systems (WE1, WE2) significantly ($p<0.05$) outperform the emphatic reference (Ref1) and WE2 significantly ($p<0.05$) outperforms WE1 in terms of emphasis effectiveness.

Table 2. MOS results for word emphasis with μ±95% confidence interval and p-value against Ref0

MOS	Ref0	Ref1	WE1	WE2
Emph.	2.29± 0.07	**2.92±0.07**, p<0.01	**3.25 ± 0.08**, p<0.01	**3.55 ± 0.07**, p<0.01
Quality	3.49 ± 0.05	**3.56±0.06,** p=0.049	**3.57±0.05,** p=0.017	3.42±0.06

For sentence emphasis, we tested three following systems: the neutral prosody reference model, *Ref0*, the WE1 word emphasis model with lower pitch parameter of α=0.1 (to reduce potential quality degradation) that was applied on all the meaningful words (i.e. the words with the two highest levels of the input rule-based word prominence), *Ref-WE*, and the proposed emphatic sentence model, *SE*. As in the word emphasis test, the subjects were given the standard quality and naturalness MOS test and were also asked to assess the emphasis, this time for the whole sentence and not for specific words. The evaluation scores are reported along with their 95% confidence interval in Table 3. The bold results are statistically significant compared to the reference system (Ref0). It can be seen that the *emphatic sentence* model significantly improves the emphasis effectiveness while preserving the quality. As we can see, applying the word emphasis model on roughly all the meaningful words in a sentence results in significant quality degradation, even when using a weak emphasis, while achieving no emphasis improvement at the sentence level. The samples for listening are available online [12].

Table 3. MOS results for sentence emphasis with μ±95% conf. and p-value against Ref0

MOS	Ref0	Ref-WE	SE
Emph.	3.12 ± 0.06	3.13 ± 0.06	**3.34 ± 0.06,** p<0.01
quality	3.67 ± 0.05	**3.43 ± 0.05**, p<0.01	3.65 ± 0.05

6. SUMMARY AND FUTURE WORK

In this work we presented a flexible emphatic prosody generation model based on Deep Bidirectional LSTM for controllable word-level emphasis realization. The word emphasis DRNN model was trained on syllable-level piece-wise linear prosodic trajectory parameters. A special data preprocessing technique was introduced to enable emphasis strength control, allowing to generate convincing emphatic prosody with no quality degradation. Additionally, we trained a BiRNN-LSTM model for emphatic sentence prosody prediction. Subjective experiments demonstrated that the synthesized speech based on this model indeed was perceived as empathic, while preserving quality and naturalness of the original. The next step of the research is to explore multi-voice training of the proposed emphatic models and their application to unseen voices and languages.

REFERENCES

[1] V. Strom, A. Nenkova, R. AJ Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky: "Modeling prominence and emphasis improves unit-selection synthesis." In Proceedings of Interspeech, 2007.

[2] K. Yu, F. Mairesse and S. Young: "Word-level Emphasis Modelling in HMM-based Speech Synthesis." Proc. ICASSP-2010, Dallas, TX

[3] A. Raux and A. W. Black: "A unit selection approach to F0 modeling and its application to emphasis," 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721), 2003, pp. 700-705.

[4] F. Meng, Z. Wu, H. M. Meng, J. Jia, L. Cai: "Hierarchical English Emphatic Speech Synthesis Based on HMM with Limited Training Data." In Interspeech, 2012.

[5] Q. T. Do, T. Toda, G. Neubig, S. Sakti, S. Nakamura: "A Hybrid System for Continuous Word-Level Emphasis Modeling Based on HMM State Clustering and Adaptive Training." In Interspeech, 2016.

[6] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson and P. Szczepaniak: "Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices." In Interspeech, 2016.

[7] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu: "WaveNet: A generative model for raw audio." arXiv:1609.03499, 2016.

[8] R. Fernandez, A. Rendel, B. Ramabhadran and R. Hoory: "Using deep bidirectional recurrent neural networks for prosodic-target prediction in a unit-selection text-to-speech system." In Interspeech, 2015.

[9] P. Mertens: "The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model." in *Proceedings of Speech Prosody, 2004*.

[10] H. Tang, X. Zhou, M. Odisio, M. Hasegawa-Johnson and T. Huang: "Two-Stage prosody prediction for emotional text-to-speech synthesis." Proc. Interspeech 2008, pp.2138-2141.

[11] F. Ribeiro, D. Florêncio, C. Zhang and M. Seltzer: "crowdMOS: An Approach for Crowdsourcing Mean Opinion Score Studies," in Proc. IEEE ICASSP, 2011

[12] S. Shechtman and M. Mordechay: Media files for "Emphatic Speech Prosody Prediction with Deep LSTM Networks." https://resedit.watson.ibm.com/researcher/view_person_subpage.php?id=9272