INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Neural Network-Based Modeling of Phonetic Durations

Xizi Wei, Melvyn Hunt, Adrian Skilling

Apple Inc

xxw395@bham.ac.uk, {Melvyn_Hunt,askilling}@apple.com

Abstract

A deep neural network (DNN)-based model has been developed to predict non-parametric distributions of durations of phonemes in specified phonetic contexts and used to explore which factors influence durations most. Major factors in US English are pre-pausal lengthening, lexical stress, and speaking rate. The model can be used to check that text-to-speech (TTS) training speech follows the script and words are pronounced as expected. Duration prediction is poorer with training speech for automatic speech recognition (ASR) because the training corpus typically consists of single utterances from many speakers and is often noisy or casually spoken. Low probability durations in ASR training material nevertheless mostly correspond to non-standard speech, with some having disfluencies. Children's speech is disproportionately present in these utterances, since children show much more variation in timing.

Index Terms: Duration Modeling, Deep Neural Networks, Phonetic Features, Lexical Stress and Pre-pausal Lengthening, TTS, ASR.

1. Introduction

Much of the past work on phonetic duration falls into three categories, aimed at gaining phonetic insight, improving the quality of TTS and improving the accuracy of ASR. In the first category, researchers have examined the extent to which certain phonetic factors have an influence on duration (*e.g.* lexical stress [1, 2], pre-pausal lengthening [1, 3], position [4], word predictability [5, 6, 7] and speaking rate [8]). Typically, only a single factor is studied at a time, and the amount of speech data is small and is taken from just one speaker or a small number of speakers (30 or fewer). Some interesting linguistic questions have been investigated in this way [5, 6, 7]. In the second category, durations are modeled parametrically or non-parametrically using DNNs or LSTM-RNNs trained on much more data than in the first category to set the durations at runtime in a parametric speech synthesizer [9, 10, 11, 12]. Typically, hundreds of phonetic features are included and there is no attempt to study the influence of any of these features. The third category is aimed at improving ASR accuracy by attempting to improve the weak duration modeling provided by standard HMM's using so-called Hidden Semi-Markov Models [13, 14]. No insight is sought into the influences on duration in this category. Although some improvements in accuracy have been claimed, the methods have not been widely adopted. This third category also includes duration modeling applied to speech recognition at the whole-word level [15, 16], though this approach is effectively limited to small-vocabulary systems (specifically, digits), which are no longer widely used.

The work reported here provides some insight into the phonetic factors controlling duration and aims ultimately to help improve both speech synthesis and recognition. A DNN is used to generate non-parametric output distributions over durations given the phonetic context for each phoneme. We incorporate the duration factors in the model in three ways to investigate their effects on duration prediction individually or in group (see Section 3.2.1). From the output distributions given by the models with or without the lexical stress and pre-pausal information, we show that the DNN is able to learn the lengthening effect of these two features (Section 3.2.2). More data is used than in any other work we are aware of, both in speaker-specific investigations and in speaker-independent investigations, where data from tens of thousands of speakers is used. The most immediate application of this work is to training speech synthesis and recognition systems, where anomalous phonetic durations can indicate discrepancies between a transcription or script and what was actually spoken.

The first author is a PhD candidate at the University of Birmingham. The work was carried out during her internship at Apple UK.

2. Method

2.1. Neural Network-Based Modeling

We used a feedforward DNN to model the duration, as shown in Figure 1. The DNN comprises a stack of fully connected layers with the softmax function [17] at the output layer. We use the same number of units in each of the hidden layers. We use rectified linear activation (*ReLU*) for the hidden units and cross-entropy as the loss function. The training procedure is optimized using *ADAM* [18].

2.2. Input features

The inputs to the DNN are a concatenation of three types of information: identity of the current phoneme, phonetic properties of adjacent phonemes and duration-related features of the phonemes. The identity of the current phoneme is encoded using a one-hot vector, while the phonetic properties of adjacent phonemes are characterized in a smaller vector (typically 15-dimensions). These phonetic properties include: long/short vowel, voiced/unvoiced consonant, plosive, affricate, nasal, fricative, glide, rhotic, sonorant, labial, alveolar, velar, aspirated and flap. The duration-related features are:

- **lexical stress**: It has been widely reported that stressed syllables are usually longer than unstressed syllables [1, 2]. When stress information is available we use one bit to show whether the current phone is in a stressed syllable or not, and we also add the stress feature to the current phone and to adjacent phones, since the position relative to the stressed syllable also affects duration.
- **pre-pausal lengthening**: Speech sounds generally lengthen before a pause [1]. We add one feature to the central phone to indicate the distance between that phone and the next pause. The value = 1/n, when n, the number of phonemes to the following pause, =1,2,3,4 or 5; or 0 for n>5.
- **position in the syllable**: Draws information about position from this feature. One bit to show whether the current phone is a consonant preceding the vowel in a syllable. We add this feature to the side phones as well as to the current phone.

http://dx.doi.org/10.21437/Interspeech.2019-2102

- **word predictability (LM scores)**: Studies in Vietnamese [5], Mandarin [6] and English [7] have found that function words are spoken more quickly than non-function words and common words more quickly than rare words, suggesting that this behavior may be language-universal. These studies are consistent with the idea that words with a higher information content (*i.e.* those that are less predictable) are spoken more carefully and hence more slowly. We use n-gram language model scores (on a log probability scale) to indicate the predictability of the word as an inverse measure of its information content.
- **speaking rate**: We use the ratio of the actual duration of the utterance to the duration the utterance would have given the expected phone durations as speaking rate. The expected phone duration is the average duration of that phone across the whole dataset.
- **peak fundamental frequency (F0)**: Peak F0 in the vowel was expected to influence duration through its association with *focal lenthening* [19] in an utterance. However, we were unable to find any influence of peak F0 on duration and it will not be discussed further.

2.3. Outputs

We obtain the reference durations from forced-alignment. We group them into 45 bins, starting at a bin corresponding to 30 ms, which is the shortest possible duration (one frame per state with three-state acoustic models) and increasing by 10 ms (one frame) for the first 39 bins. Beyond that point, there are too few samples at 10 ms spacing and the bins are made progressively wider. For example, the 40th and 41st bins correspond to 20ms and 30ms spacing respectively. Durations larger than 670ms are all put into the 45th bin.

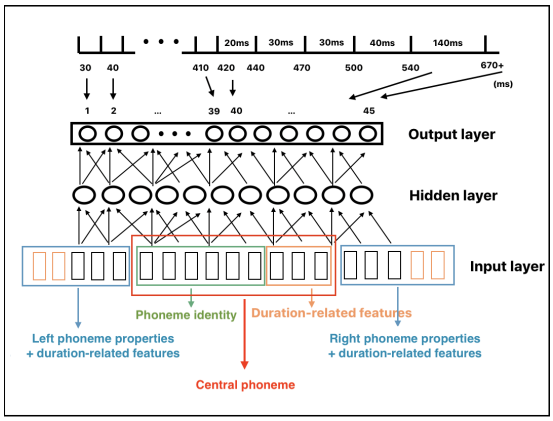


Figure 1: *An overview of the duration modeling.*

2.4. Outliers

We detect outliers using the output from the DNN. We use the value of the bin to which the reference duration belongs as the probability of the duration, and by ranking the probabilities we can get a list of phonemes with the lowest probabilities. These phonemes with unlikely duration, which we regard as outliers, can indicate misalignments or departures from the transcription. Examples are shown in Section 4.1.

3. Experiments and Results

We measure the basic effectiveness of our modeling in three ways: (i) cross-entropy loss on a test set, (ii) a measure we call "*precision*", which is the proportion of measured durations whose bin exactly matches the mode of the model's predicted duration distribution. For most predictions (*i.e.* those below 410ms), this is within 10 ms (one frame), and thus the highest precision possible, and (iii) a precision with more tolerance that counts not only the match to the bin corresponding to the peak of the distribution but also the neighboring bin on each side. We denote these three measurements as *CE_loss*, *precision* and *precision_3* respectively. All neural networks are built using *Pytorch* [20].

3.1. Data

We have in-house datasets from two native speakers of American English recorded for TTS purposes. One of the speakers SPK1, is female and the other, SPK2, male. There are 64,795 utterances (33 hours) in the SPK1 dataset and 27,550 utterances (13 hours) in the SPK2 dataset. We also have an in-house dataset, SPK-ASR, of less carefully controlled recordings intended for ASR that contains 540,389 utterances from 535,556 speakers of all ages, including children. We used forced alignment to get the duration for each phone. We used SPK1 and SPK2 for speaker-dependent modeling and SPK-ASR for speaker-independent modeling. The phonetic symbol sets used in these three datasets are different. We used 46, 42 and 50 one-hot vectors to encode the central phone identity for the three datasets respectively.

3.2. Speaker-dependent modeling

3.2.1. Duration-related features configurations

We used a DNN with 2 hidden layers and 256 hidden units in each layer as a baseline for exploring the feature configurations. We used a minibatch of 64 and train the model for 30 epochs. The learning rate was 0.001 for each epoch. We found that the final result depended to some extent on the random start point of the model built. We therefore ran our precision tests 10 times, each with a different random start and a different randomly selected test set. For each test, we randomly sampled the dataset to use 90% for training and 10% for testing. We then computed the overall mean and standard error of the precision. We began with *Baseline_0* trained on SPK1 (the input to the neural network being just the one-hot vector that encodes the identity of the current phone with no context) and obtained a precision of 19%. In *Baseline_1* the input has a context of ± 1 (1 phone on each side of the current phone) and the precision increased to 28.68%. Thus, we obtained a 9.28% absolute increase by including context.

Baseline_1 was then augmented with the duration-related features in three ways: (i) adding each one to the Baseline_1 to see the effect on its own, (ii) cumulatively adding the features to the Baseline_1 and (iii) including all the features except the named one. The results are shown in Table 1. The precision increases as the duration-related features are added, among which the stress has the biggest positive effect. The speaking rate for the utterance and pre-pausal lengthening also have a strong influence. The location of consonants within a syllable (*i.e.* before or after the vowel) has a somewhat weaker influence, as has the predictability of the word containing the phoneme as estimated by a stochastic language model. We also carried out the cumulative experiments on another TTS dataset, SPK2. Figure 2 shows that the effect of these factors is very similar for a different speaker.

Table 1: *The duration prediction results for Baseline_1 with different feature configurations trained on SPK1. The standard errors of the precisions (%) are in the range from 0.0002 to 0.006.*

Models	(i) Just the named feature			(ii) Cumulative			(iii) Leave one out		
	precision	precision_3	CE_loss	precision	precision_3	CE_loss	precision	precision_3	CE_loss
pre/post-vocalic	29.23	66.08	0.0296	29.23	66.08	0.0296	32.76	72.30	0.0273
stress	30.63	68.47	0.0287	30.76	68.62	0.0286	31.42	69.87	0.0282
pre-pausal	30.16	67.93	0.0291	32.28	71.26	0.0277	31.81	70.49	0.0279
predictability	29.19	66.08	0.0298	32.52	71.70	0.0276	32.74	72.28	0.0272
speaking rate	29.54	66.59	0.0294	33.08	72.73	0.0272	32.52	71.70	0.0276

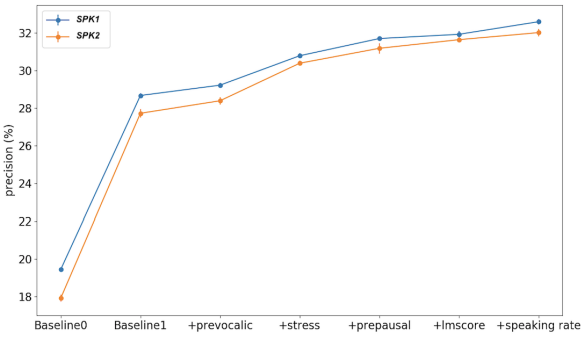


Figure 2: *Models with different feature configurations trained separately on two speakers: SPK1 and SPK2.*

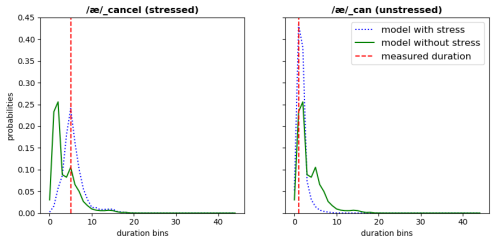
3.2.2. *Stress and pre-pausal lengthening effect*

The duration probability distributions in Figure 3 give examples showing that the network is able to learn the lengthening effect of stress and pre-pausal features, and the predictions are closer to the measured duration bins (red dashed lines) with these two features on. The /æ/ in "cancel" and "can", which we denote as "æ_cancel" and "æ_can", have the same context, but different stress values ("can" as a modal verb normally being unstressed). In Figure 3a, the two green curves are the same, but knowledge of stress increases the predicted duration for /æ/ in "cancel" and reduces it in "can". Figure 3b compares distributions with and without an input providing the distance to the next pause. Predicted duration distributions for utterance-final "here" are shown left to right as the three phonemes /h/ /i/ /ɝ/. Since the baseline model here has a context of ± 1, the distribution of /ɝ/ in "here" still has the effect of pre-pausal lengthening even without that feature given in the input features. For the /h/ and /i/, knowledge of pause proximity increases the predicted duration.

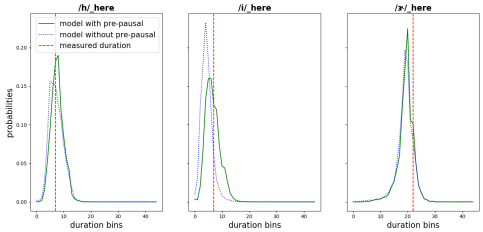
We evaluated the SPK2-trained model on the SPK1 testing set and the SPK1-trained model on the SPK2 test set. Since SPK1 has much more training data than SPK2, we also evaluated the SPK1-trained model with reduced training data size. The results shown in Table 2 suggest that the precision decreases considerably when testing on a different speaker. The results in row 3 with a model trained on SPK1 when using a reduced set to match that available for SPK2 match much more closely the results from training on SPK2 (row 1), suggesting that the difference between the results with the two speakers is largely attributable to the discrepancy in the amount of training material and indicating that more than 10 hours of training speech is needed for optimal model training. This result also largely explains the offset between the two curves in Figure 2.

3.2.3. *Model configurations*

We trained the DNN with a range of hidden layers ($d \in 1, 2, 3$) and hidden units in each layer ($w \in 128, 256, 512$) and wider



(a) */æ/ in "cancel" and "can" comparing model outputs when lexical stress information is or is not included.*



(b) */h/ /i/ /ɝ/ in an utterance-final "here", comparing distributions with and without an input providing the distance to the next pause.*

Figure 3: *Predicted duration distributions. The red dotted line shows the measured duration for one example.*

Table 2: *Cross-speaker precision tests (%). The models used all the duration-related features.*

training	SPK2_test (1h)	SPK1_test (3h)
SPK2 (10h)	31.00	22.70
SPK1 (30h)	23.34	32.56
SPK1 (10h)	22.84	31.45

context (± 1, ± 2 and ± 3). As shown in Figure 4, the precision improves as the number of parameters in the model is increased and the context is longer. Taking computational efficiency into account, the best configuration for now is three hidden layers with 256 hidden units in each layer and with ± 3 context, which achieves a precision of 35.67% and precision_3 of 89.88%.

3.3. Speaker-independent modeling

We applied our duration modeling method with the configuration in Section 3.2.3 to speaker-independent modeling with 80% of the SPK-ASR dataset as the training data. We obtained a precision of 10.50% and precision_3 of 40.30% on a testing set (10%). It is more challenging because in the SPK-ASR corpus almost every utterance is from a different speaker and spoken in a spontaneous way. Moreover, stress and LM score have not yet been incorporated.

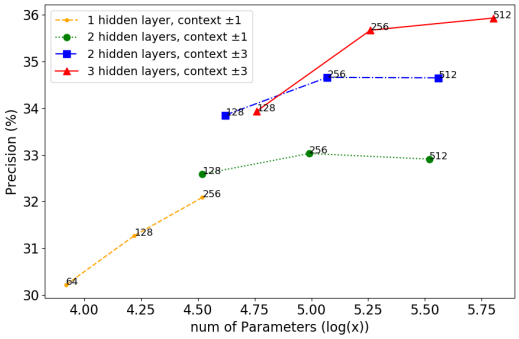


Figure 4: *Model configurations.*

4. Applications

4.1. Outlier detection in TTS and ASR

We use the best configuration from Section 3.2.3 to detect outliers for the SPK1 TTS dataset and the ASR dataset.

Figure 5a, 5b and 5c show three examples from the top outliers in the SPK1 dataset corresponding to three kinds of problems that have been seen to occur in the TTS training corpus. 48 out of 50 outliers are correctly detected as having bad alignments.

We also observed misalignments in the ASR dataset as well as disfluencies as in Figure 5d (such disfluencies being rare in the speech of the professional speakers producing the TTS dataset). When examining outliers in ASR (Table 3), 12 out of the top 50 outliers (*i.e.* 24%) were found to be from children. By contrast, just 11 out 100 randomly selected utterances were judged to be from children, suggesting that children make a disproportionate contribution to the set of outliers. Among the top 50 outliers, 8% are because of disfluencies, resulting in bad transcriptions and hence bad alignments. By contrast, in the randomly selected utterances, fewer than 2% were found to have bad alignments. The rest of the outliers evidently get their low scores because the speaker was dictating and hence speaking slowly, had put extreme stress on a word and hence lengthened it, or was speaking in a playful style.

Table 3: *Proportion of children's speech and bad alignments in the top 50 outliers and the randomly selected utterances for the ASR data.*

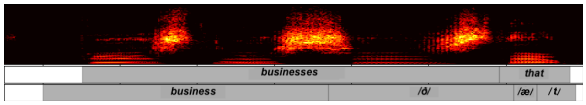
	Top 50 outliers	Random utts
Children's speech	24%	11%
Bad alignments	8%	<2%

5. Conclusions

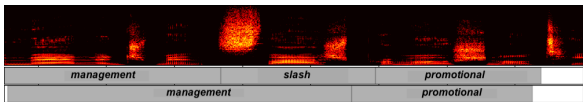
A DNN can provide a useful prediction of the distribution of durations of a phoneme in a specified context. It offers a technique for gaining a basic understanding from large speech corpora (rather than the more usual small set of examples) of how various factors combine to determine phonetic durations in a given language. The prediction is best, at least in American English, when the phonetic properties of at least three phonemes on each side of the phoneme under consideration are provided to the DNN, together with other relevant information, such as lexical stress in the syllable and estimated average speaking rate.

Distributions produced in this way can be used to spot improbable durations that often arise from a mismatch between the speech and either the words in the phonetic transcription or the dictionary pronunciations of those words. Low probability durations may also occur because the speech is particularly expressive. In training material for TTS these anomalies can be used to correct transcriptions and dictionary entries as well as to exclude unsuitable speech from the TTS training set. In ASR training material, low duration scores may result from disfluencies (rare in TTS training speech), but the most common cause from our limited sampling of the outliers appears to be unusual timing from expressive speech or dictation mode. This second cause does not invalidate the speech for ASR training purposes, though the first clearly does. Children's speech is overrepresented in the set of low duration scores because phonetic durations in their speech appear to be much more variable than those of adults' speech.

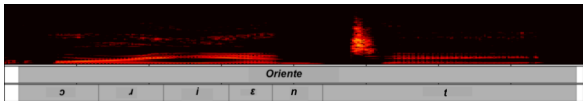
So far, this work has been confined to American English. We might speculate that duration information will be particularly useful for ASR in languages such as Japanese, Finnish, Estonian and Arabic [21] that have phonemic length.



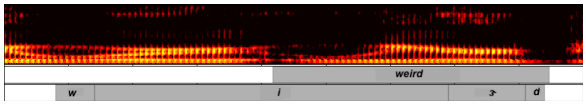
(a) *deviation from the script: the speaker says "businesses", but the transcription has "business"; the /ð/ in "that" is consequently misaligned to the end of "businesses".*



(b) *deviation from the script: the speaker says "management slash promotional", having evidently read "management/promotional", but the transcription has "management promotional", thus the /t/ is misaligned to an unlikely long duration.*



(c) *mismatch in the way the word is pronounced relative to the dictionary: the speaker says "Oriente" as /ɔɹi'ɛnte/, but the pronunciation in the dictionary for "Oriente" is /aɹi'ɛnt/ without a final vowel, causing the /t/ to be misaligned to a longer segment.*



(d) *mistranscription: the speaker actually says "wei... weird" but the transcription is simply "weird"; the /i/ is consequently aligned to a much longer portion of speech.*

Figure 5: *Outlier examples, the upper annotation line is what the speaker says and the lower is from the forced alignment.*

6. Acknowledgments

Earlier work on seeking factors that influence durations was carried out by Dominic Hunt and especially by Paul Coles. John Bridle, Barry Theobald and Rin Metcalf made many useful comments and Felipe Espic pointed us to the *ADAM* optimizer.

7. References

[1] D. H. Klatt, "Linguistic uses of segmental duration in english: acoustic and perceptual evidence," *The Journal of the Acoustical Society of America*, vol. 59, no. 5, pp. 1208–21, 1976.

[2] T. H. Crystal and A. S. House, "Segmental durations in connected-speech signals: Syllabic stress," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1574–85, 1988.

[3] W. N. Campbell and S. Isard, "Segment durations in a syllable frame," *Journal of Phonetics*, vol. 19, no. 1, pp. 37–47, 1991.

[4] P. A. Luce and J. Charles-Luce, "Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production," *The Journal of the Acoustical Society of America*, vol. 78, no. 6, pp. 1949–57, 1985.

[5] M. Brunelle, D. Chow, and T. N. U. Nguyen, "Effects of lexical frequency and lexical category on the duration of Vietnamese syllables," in *ICPhS*, 2015.

[6] E. Sherr-Ziarko, "Word frequency effects on homophonous words in Mandarin Chinese," in *ICPhS*, 2015.

[7] A. Bell, J. Brenier, M. Gregory, C. Girand, and D. Jurafsky, "Predictability effects on durations of content and function words in conversational English," *Journal of Memory and Language*, vol. 60, no. 01, pp. 92–111, 2009.

[8] E. Chodroff, J. Godfrey, S. Khudanpur, and C. Wilson, "Structured variability in acoustic realization: A corpus study of voice onset time in American English stops," in *ICPhS*, 2015.

[9] G. E. Henter, S. Ronanki, O. Watts, M. Wester, Z. Wu, and S. King, "Robust TTS duration modelling using DNNs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5130–5134.

[10] S. Ronanki, O. Watts, S. King, and G. E. Henter, "Median-based generation of synthetic speech durations using a non-parametric approach," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 686–692.

[11] B. Chen, T. Bian, and K. Yu, "Discrete duration model for speech synthesis," in *INTERSPEECH*, 2017.

[12] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[13] M. Russell and R. Moore, "Explicit modelling of state occupancy in Hidden Markov Models for automatic speech recognition," in *IEEE International Conferenceon Acoustics, Speech and Signal Processing (ICASSP)*, April 1985.

[14] K. Oura, Y. Nankaku, and K. Tokuda, "Hidden semi-Markov Model based speech recognition system using weighted finite-state transducer," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2006.

[15] N. Ma and P. D. Green, "Context-dependent word duration modelling for robust speech recognition," in *INTERSPEECH*, 2005.

[16] K. Power, "Durational modelling for improved connected digit recognition," in *ICSLP*, 1996.

[17] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing*, 1990, pp. 227–236.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 12 2014.

[19] S. J. Eady, W. E. Cooper, G. V. Klouda, P. R. Mueller, and D. W. Lotts, "Acoustical characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments," *Language and Speech*, vol. 29 3, pp. 233–251, 1986.

[20] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS-W*, 2017.

[21] I. Zangar, Z. Mnasri, V. Colotte, D. Jouvet, and A. Houidhek, "Duration modeling using DNN for Arabic speech synthesis," in *9th International Conference on Speech Prosody*, Poznań, Poland, June 2018.