Continuous Fundamental Frequency Prediction with Deep Neural Networks

Bálint Pál Tóth, Tamás Gábor Csapó
Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, HUNGARY
{toth.b, csapot}@tmit.bme.hu

Abstract— **Deep learning is proven to outperform other machine learning methods in numerous research fields. However, previous approaches, like multispace probability distribution hidden Markov models still surpass deep learning methods in the prediction accuracy of speech fundamental frequency (F0), inter alia, due to its discontinuous behavior. The current research focuses on the application of feedforward deep neural networks (DNNs) for modeling continuous F0 extracted by a recent vocoding technique. In order to achieve lower validation error, hyperparameter optimization with manual grid search was carried out. The results of objective and subjective evaluations show that using continuous F0 trajectories, DNNs can reach the modeling performance of previous state-of-the-art solutions. The complexity of DNN architectures could be reduced in case of continuous F0 contours as well.**

Keywords-feedforward deep neural networks, speech synthesis, fundamental frequency, F0

I. INTRODUCTION

Due to the revolutionary increase in the amount of available data, the rise of high performance GPUs and the novel results in neural networks, deep learning has received high attention among machine learning and speech scientists. The numerous layers of deep architectures are able to extract different abstractions of the input data and predict or classify them efficiently.

The history of neural networks in speech research has started in the 90s [1]–[3]. The prediction of speech fundamental frequency (F0) with neural networks - that is the topic of the current paper - has had promising results even about 25 years ago [4]. However, due to the lack of the recent progress of technology and new machine learning algorithms, neural networks were unable to vanquish the state-of-the-art solutions of that time. In speech research, after the data driven unit selection era, the statistical parametric speech synthesis, mostly hidden Markov-model based text-to-speech synthesis (HMM-TTS) gained a lot of interest [5]. In HMM-TTS very-rich contexts are modeled by decision tree-clustered context-dependent HMMs. Nevertheless, decision trees are not suitable to model complex, many-to-many dependencies. Furthermore, despite their advantages, the Gaussians underlying the context-dependent HMMs are inefficient to model data that lie on or near a nonlinear manifold in the data space. Speech is considered to have such a behavior according to Hinton and his colleagues [6]. Deep neural networks (DNNs) can overcome both limitations and can solve arbitrary non-linear problems if enough units in the hidden layers and sufficient amount of training data is available.

Modeling speech generation with deep neural networks has significant results, yet there is much space for improvements. Zen and his colleagues [7] were among the firsts who created a feedforward DNN speech synthesis system. It was able to approach the spectral modeling capability of HMM-TTS systems, but it produced worse results for F0 trajectories. Similar phenomena occurred with other neural network architectures, like deep belief nets [8], [9] and bidirectional long short-term memory [10]. There have been investigations with promising results on training deep neural networks with different approaches than the traditional pulse-noise vocoder, for example GlottHMM [11], [12] and STRAIGHT [13]. The authors of the current paper share the view that today's speech representation (extracting a large number of parameters in every 25 ms) is still far from the underlying dynamic parameters of the human speech production system. Therefore, in the current research, we are investigating a simpler representation of speech production that is easier to predict - namely a continuous F0 model.

Traditionally, using standard pitch tracking methods in vocoders, the F0 contour is discontinuous at voiced-unvoiced (V-UV) and unvoiced-voiced (UV-V) boundaries, because F0 is not defined in unvoiced sounds. This can pose several issues in statistical modeling. For example, in HMM-TTS, Multi-Space Distribution (MSD) was proposed for discontinuous F0 modeling, which involves building separate models for voiced and unvoiced frames of speech [14]. However, it has been recently shown that excitation models using continuous F0 have several advantages in statistical parametric speech synthesis [15]. First of all, using a continuous F0 contour, the ineffective MSD-HMM modeling around V-UV and UV-V transitions can be omitted. Second, it was found that more expressive F0 contours can be generated using a continuous F0 than using the standard discontinuous F0 models [16]. In such continuous systems, often a separate stream of voicing strength or label is used for modeling the voicing feature [17]. Furthermore, the voiced/unvoiced (V/UV) decision can be left up to the aperiodicity features in a mixed excitation vocoder [18] or to the dynamic voiced frequency feature in a residual-based vocoder [19], [20]. In [21], an excitation model has been proposed which combines continuous F0 modeling with Maximum Voiced Frequency (MVF). This model has been shown to produce more natural synthesized speech for voiced

sounds than traditional vocoders based on standard pitch tracking, whereas it was also found that there is room for improvement in modeling unvoiced sounds with this vocoder.

The authors' purpose is to investigate the modeling capability of deep neural networks and the model complexity of F0 trajectories extracted by traditional (discontinuous) and continuous vocoders. Our hypothesis is that the perceptual quality of DNN-based prediction using continuous F0 will be superior to those that use discontinuous F0.

II. METHODS

A. Baseline vocoder

For the baseline system we used a traditional pulse-noise vocoder [22] in which the fundamental frequency was extracted by the SWIPE algorithm [23]. For modeling the spectrum, 24-order Mel-Generalized Cepstral (MGC) [24] analysis is performed on the speech signal with α=0.42 and γ=−⅓ parameters. In this kind of vocoder, the F0 is separated into voiced and unvoiced regions as pointed out in Section I. The excitation of voiced regions consists of series of impulses, while unvoiced regions have noise type excitation. The vocoder stores for every window (25 ms long, 5 ms shift) a voiced/unvoiced flag (V/UV flag) and the actual F0 value for voiced regions. The discontinuity of such a method increases the complexity of the data space. Therefore, in case of DNN trainings, we interpolated the unvoiced regions linearly and train the neural networks with the interpolated F0 (that is continuous) and with the V/UV flag. The models trained by the baseline vocoder are referred to as F0std.

B. Vocoder with continuous F0

For vocoding with continuous F0, we use a recent vocoder [21]. From the input speech waveform sampled at 16 kHz, MGC analysis is performed with the same parameters as in the baseline vocoder. Fundamental frequency is calculated by the open source implementation of a simple continuous pitch tracker[1] [15] denoted as F0cont. In regions of unvoiced sounds, this pitch tracker interpolates F0 based on a linear dynamic system and Kalman smoothing. After this step, Maximum Voiced Frequency is estimated from the speech signal using the MVF_Toolkit[2] [19] , resulting in the MVF parameter stream. In all steps, 5 ms frame shift is used.

To synthesize voiced excitation, we are using principal component analysis (PCA) based residual frames, as they have been shown to overcome simple impulse based excitation [20]. First, PCA residuals are overlap-added depending on the F0cont parameter, resulting in the voiced component of the excitation. The unvoiced component of the excitation is based on white noise. As there is no strict voiced/unvoiced decision or parameter stream in this vocoder, the MVF parameter models the voicing information: for unvoiced sounds, the MVF is low (around 1 kHz), for voiced sounds, the MVF is high (typically above 4 kHz), whereas for mixed excitation sounds, the MVF is in between. In a frame-by-frame basis, voiced excitation is low pass filtered corresponding to MVF, while unvoiced excitation is high pass filtered based on the MVF value, and these two frequency components are added together. Finally, the speech is reconstructed from the excitation and the MGC parameter stream using an MGLSA (Mel-Generalized Log Spectrum Approximation) filter [25]. The models trained by this vocoder are referred to as F0cont.

[1] https://github.com/idiap/ssp

[2] http://tcts.fpms.ac.be/~drugman/files/MVF.zip

C. Training of hidden Markov models

We wanted to compare the modeling capacity of HMM and DNN statistical methods for the task of both traditional and continuous F0 prediction. For training the F0 contours with HMMs, the standard HTS toolkit is used [22]. In case of F0std, multi space distribution training is applied [14], whereas for F0cont, we use simple HMMs. The first and second derivatives of the parameters are also stored in the parameter files and used in the training and generation phases. Decision tree-based context clustering is used with context dependent labeling applied in the Hungarian version of HTS 2.3beta [22], [26]. Independent decision trees are built for all the parameters and duration using a maximum likelihood criterion. Although durations and MGC were also trained in the system, in the evaluation part of this paper only the modeled F0 stream is used combined with other parameters obtained from the natural sentences.

D. Training of deep neural networks

In this research we focused on feedforward deep neural networks. The output and input features are introduced in Table I and II and the general architecture of the network is shown on Figure 1. In the training we used the squared error loss function over minibatches. For optimization we chose ADADELTA [27] because its robustness (adaptive learning rate control, can handle noisy gradients and different data representations). To be able to discard the computational overhead of pretraining, we used rectified linear units [28] (ReLU) as activation function in the hidden layers. After preliminary experiments we changed them to parametric rectified linear units [29] (PReLU). PReLUs only slightly increase the complexity and are able to achieve better error rates by adaptively learning the shapes of activation functions. In the output layer sigmoid was used as activation function. Xavier's weight initialization technique was used in the case of input-hidden and hidden-output weights [30]. We used orthogonal initialization between the hidden layers with zero bias. In the training dropout with 50% probability was applied after each layer except the output. Early stopping was applied - if the validation error did not decrease in 50 epochs, the training had been stopped. Both input and output features were transformed. The input features were standardized to have zero mean and unit variance. The output features were normalized between 0.01 and 0.99 [31]. The training samples were randomly shuffled.

TABLE I. OUTPUTS OF THE NEURAL NETWORK FOR F0STD AND F0CONT VOCODERS.

System	Feature name	#	Type
F0std	LogF0	1	Continuous (interpolated)
	V/UV flag	1	Binary
F0cont	LogF0	1	Continuous
	MVF	1	Continuous

TABLE II. INPUTS OF THE NEURAL NETWORK.

Feature name	#	Type
Quinphone	5*68	One-hot
Number of phonemes/syllables/words/phrases in the previous/current/next syllable/word/phrase/sentence	4*3	Numerical
Number of syllables/words in the current sentence	2	Numerical
Forward/backward position of the actual phoneme/syllable/ word/phrase in the syllable/word/phrase/sentence	2*3	Numerical
Phone boundaries	2	Numerical
Percentual position of the actual frame within the phone	1	Numerical
Altogether:	**363**	

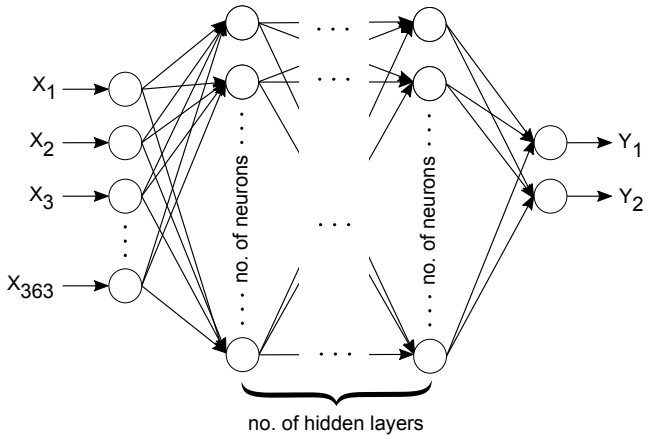

Figure 1. General architecture of the applied deep neural network. The Y_1 and Y_2 outputs are F0 and V/UV flag or F0 and MVF, in case of baseline and continuous F0, respectively.

III. EVALUATION

In the evaluation part DNNs trained with the baseline and with the continuous F0 model took place. For training, one male and one female speaker was selected from the Precisely Labelled Hungarian Database (PLHD) containing 1984 sentences [32]. Precise labelling covers manually corrected phonetic transcription and phone boundaries. Only declarative sentences were investigated. With the male speaker, an objective evaluation was carried out first to optimize hyperparameters. Based on this objective evaluation, the top five systems were trained with the female speaker's corpus as well. The training, validation and test data were the 80, 15 and 5 percentage of the corpus, respectively. The male and the female speakers were trained separately. In the evaluation part phone durations from natural utterances were used for the temporal information of the input vector. The deep neural network introduced in Section II.D was implemented in Torch7 deep learning framework [33], and the calculations were done on high performance NVidia GPUs.

A. *Objective evaluation*

We performed hyperparameter optimization with manual grid search. The hyperparameters introduced in Section II.D throughout remained the same and the following hyperparameters were investigated: number of hidden layers, number of neurons in hidden layers and size of the minibatches. Altogether 64 trainings were done for the baseline (F0std) system and 73 for F0cont.

TABLE III. THE WINNING 5-5 DEEP NEURAL NETWORK ARCHITECTURES OF HYPERPARAMETER OPTIMIZATION WITH MANUAL GRID SEARCH.

(A) BASELINE (MINIBATCH SIZE=128)

ID	# Hidden Layers	# Neurons	Epochs	Validation MSE
F0std-1	3	350	61	0.01076
F0std -2	3	650	32	0.01078
F0std -3	3	900	30	0.01089
F0std -4	3	950	36	0.01099
F0std -5	3	800	37	0.01103

(B) CONTINUOUS F0 (MINIBATCH SIZE=8)

ID	# Hidden Layers	# Neurons	Epochs	Validation MSE
F0cont-1	3	160	2	0.00239
F0cont-2	3	80	67	0.00346
F0cont-3	1	128	2	0.00349
F0cont-4	3	70	12	0.00352
F0cont-5	2	100	28	0.00356

In the hyperparameter optimization phase the number of hidden layers, the number of neurons and the size of the minibatch were set between 1..7, 80..2048, 8..256, respectively. The 5-5 best combinations of hyperparameters and the corresponding mean square errors on the validation set are shown in Table III. For the later analyses, we chose the F0std-1 and F0cont-1 hyperparameter sets.

After the hyperparameter optimization, we measured the correlation and RMSE between the F0 curves of the natural sentences and those of obtained by the two statistical methods (HMM and DNN) combined with the two F0 modeling methods (F0std and F0cont).

This calculation was done on the 5% test data for both speakers. Correlation and RMSE were only measured on the voiced frames (based on the manually labelled phonetic boundaries of the natural sentences). For calculating the correlation, we used the equation proposed by Hermes for comparing F0 contours [34]:

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

where x and y are two F0 contours ($\bar{x}$ and $\bar{y}$ denote their means). RMSE was calculated in a standard way.

The mean correlation values are shown in Figure 2. In general, the DNN statistical method resulted in lower correlation values than the HMM method for both speakers. An important result clearly visible in the figure is that F0cont has significantly higher correlation than F0std in all cases ($p<0.05$). Figure 3 shows the mean RMSE values for all combinations. For F0std, the DNN has higher errors than the HMM method, while DNN and HMM are close to each other for F0cont. The tendency between F0std and F0cont is similar to the results of correlation: here, F0cont resulted in significantly lower RMSE errors compared to F0std in all cases.

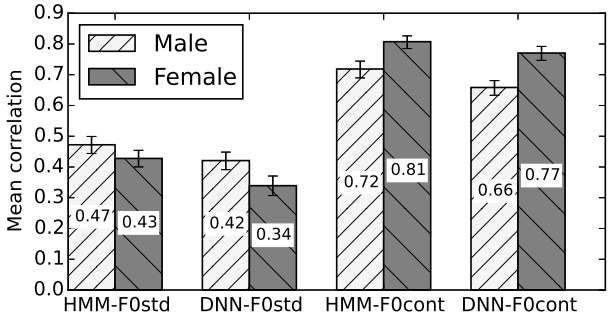

Figure 2. Mean correlation between natural F0 and modeled F0 contours. Higher value means larger similarity between the compared F0 trajectories. Error bars show the bootstrapped 95% confidence intervals.

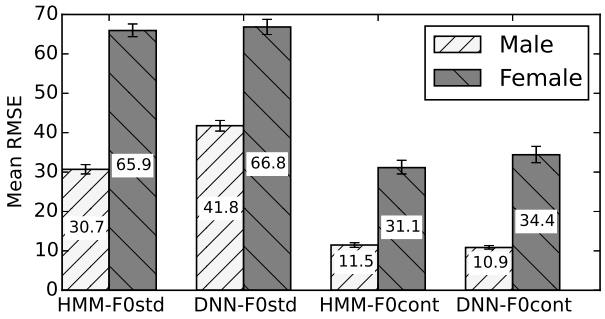

Figure 3. Mean RMSE between natural F0 and modeled F0 contours. Higher value means larger average difference between the F0 trajectories.Error bars show the bootstrapped 95% confidence intervals.

B. *Subjective evaluation*

In order to evaluate which F0 modeling method is closer to the pitch contour of natural speech, a web-based MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchor) listening test [35] was carried out. The advantage of MUSHRA is that it allows evaluating multiple samples in a single trial without breaking the task into many pairwise comparisons. Our aim was to measure the perceived intonation of sentences; therefore we compared a reference natural sentence with vocoded sentences. In all vocoded sentences, the excitation was either the result of the statistical (DNN/HMM) methods or natural sentences (NAT) in combination with the two F0 extraction methods (F0std/F0cont). The spectral filtering was always using the original MGC parameter stream of the natural utterances. In case of the F0cont vocoder, the MVF parameter was also the result of the statistical methods. We added a benchmark utterance to help the listeners to scale the other utterances. The benchmark vocoded sentences had zero F0 at the whole duration, resulting in a whispered-like speech signal. From the sentences used in the objective evaluation, the 5 sentences with the highest average RMSE were selected; these are considered the worse sentences analytically, according to Section III.A. Altogether, 80 utterances were included in the test (2 speakers × 8 types × 5 sentences). Before the test, listeners were asked to listen to an example from the male speaker to adjust the volume. In the test, the listeners had to rate the naturalness of each stimulus relative to the reference (which was the natural sentence), from 0 (highly unnatural) to 100 (highly natural). The utterances were presented in a randomized order (different for each participant).

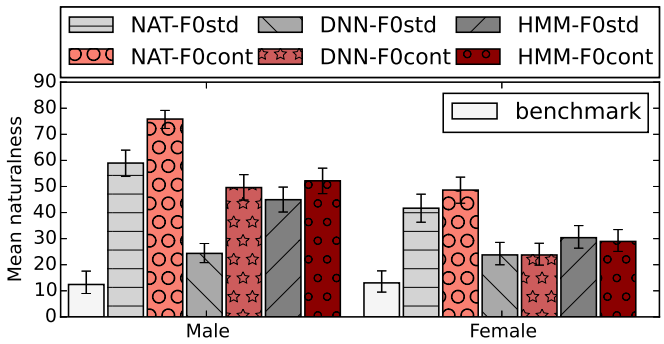

Figure 4. Results of the subjective evaluation for the naturalness question. Higher value means larger naturalness. Errorbars show the bootstrapped 95% confidence intervals. The score for the natural speech is not included, because it is always 100.

Altogether 18 listeners participated in the test (9 females, 9 males). All subjects were native Hungarian speakers, between 21-74 years (mean: 39 years). On average the test took 13 minutes to complete. The MUSHRA scores of the listening test are presented in Figure 4. for the two speakers and seven types.

The results show that the F0cont method always outperforms the F0std method for the male speaker, while F0std and F0cont are roughly equal for the female speaker (omitting the versions with the natural F0). For both speakers, the DNN statistical modeling achieved similar scores than the HMM modeling (except for the case of F0std and the male speaker). The ratings of the listeners were compared by Mann-Whitney-Wilcoxon ranksum tests as well, with a 95% confidence level, showing that there were significant differences. For the male voice, DNN-F0std was significantly less preferred than DNN-F0cont, HMM-F0std and HMM-F0cont. For the female voice, the HMM-based versions were significantly preferred over the DNN-based versions.

IV. Conclusions and Discussion

From the objective and subjective analysis, we can conclude that F0cont curves can be approximated better than F0std curves, using both HMM and DNN statistical methods. Simpler DNN models were enough for the F0cont in contrast with the models of F0std. These smaller models can be beneficial in embedded systems with limited computational resources. The convergence of F0cont models was also faster – the topmost F0cont model achieved its lowest validation error approximately 7 times faster than the best F0std model. These results suggest that the continuous representation of fundamental frequency forms a less complex system than the V/UV based F0std.

In the case of F0cont the modeling capacity of the deep neural network approaches the performance of the state-of-the-art MSD-HMM based fundamental frequency prediction. Taking into consideration that DNNs are proven to be more efficient in spectral component prediction, this result may raise the quality of feedforward DNN based speech synthesis over the HMMs.

Furthermore it must be noted that in the HMM system the first and second derivatives of F0 were used in training and

generation, and global variance was also applied [36]. Introducing dynamic features in the DNNs are expected to result in better predictions.

Both questions raised above are planned to be addressed in our further research as well as applying different neural network architectures, like long short-term memory and auto-encoder.

ACKNOWLEDGMENT

Bálint Pál Tóth gratefully acknowledges the support of NVIDIA Corporation with the donation of an NVidia Titan X GPU used for his research. We would like to thank the listeners for participating in the subjective test. We thank Philip N. Garner for providing the open source continuous pitch tracker. This research is partially supported by the Swiss National Science Foundation via the joint research project (SCOPES scheme) SP2: SCOPES project on speech prosody (SNSF no IZ73Z0_152495-1) and by the Hungarian Scientific Research Fund (OTKA) under contract ID PD-112598, "Automatic Phonological Phrase and Prosodic Event Detection for the Extraction of Syntactic and Semantic/Pragmatic Information from Speech".

REFERENCES

[1] R. P. Lippmann, "Review of neural networks for speech recognition," *Neural Comput.*, vol. 1, no. 1, pp. 1–38, Mar. 1989.

[2] M. Riedi, "A neural-network-based model of segmental duration for speech synthesis," in *Proc. Eurospeech*, 1995, pp. 599–602.

[3] T. J. Sejnowski and C. R. Rosenberg, "Parallel networks that learn to pronounce english text," *Complex Syst.*, vol. 1, pp. 145–168, 1987.

[4] C. Traber, "F0 generation with a data base of natural F0 patterns and with a neural network," in *Proc. ISCA SSW1*, 1990, pp. 141–144.

[5] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[6] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[7] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[8] S. Kang, X. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," in *Proc. ICASSP*, 2013, pp. 8012–8016.

[9] S. Mukherjee, "F0 modeling in HMM-based speech synthesis system using Deep Belief Network," in *COCOSDA*, 2014, pp. 1–5.

[10] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. Interspeech*, 2014, pp. 1964–1968.

[11] T. Raitio, H. Lu, J. Kane, A. Suni, M. Vainio, S. King, and P. Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *Proc. EUSIPCO*, 2014, pp. 2290–2294.

[12] T. Raitio, A. Suni, L. Juvela, M. Vainio, and P. Alku, "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort," in *Proc. Interspeech*, 2014, pp. 1969–1973.

[13] S. Takaki, S. Kim, J. Yamagishi, and J. Kim, "Multiple feed-forward deep neural networks for statistical parametric speech synthesis," in *Proc. Interspeech*, 2015, pp. 2242–2246.

[14] K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[15] P. N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," *IEEE Signal Process. Lett.*, vol. 20, no. 1, pp. 102–105, 2013.

[16] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 5, pp. 1071–1079, 2011.

[17] Q. Zhang, F. K. Soong, Y. Qian, Z. Yan, J. Pan, and Y. Yan, "Improved modeling for F0 generation and V/U decision in HMM-based TTS," in *Proc. ICASSP*, 2010, pp. 4606–4609.

[18] J. Latorre, M. J. F. Gales, S. Buchholz, K. Knil, M. Tamura, Y. Ohtani, and M. Akamine, "Continuous F0 in the source-excitation generation for HMM-based TTS: Do we need voiced/unvoiced classification?," in *Proc. ICASSP*, 2011, pp. 4724–4727.

[19] T. Drugman and Y. Stylianou, "Maximum voiced frequency estimation: exploiting amplitude and phase spectra," *IEEE Signal Process. Lett.*, vol. 21, no. 10, pp. 1230–1234, 2014.

[20] T. Drugman and T. Raitio, "Excitation modeling for HMM-based speech synthesis: breaking down the impact of periodic and aperiodic components," in *Proc. ICASSP*, 2014, pp. 260–264.

[21] T. G. Csapó, G. Németh, and M. Cernak, "Residual-based excitation with continuous F0 modeling in HMM-based speech synthesis," in *Lecture Notes in Artificial Intelligence*, vol. 9449, A.-H. Dediu, C. Martín-Vide, and K. Vicsi, Eds. Budapest, Hungary: Springer International Publishing, 2015, pp. 27–38.

[22] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, and A. W. Black, "The HMM-based speech synthesis system version 2.0," in *Proc. ISCA SSW6*, 2007, pp. 294–299.

[23] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, Sep. 2008.

[24] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *Proc. ICSLP*, 1994, pp. 1043–1046.

[25] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electron. Commun. Japan (Part I Commun.*, vol. 66, no. 2, pp. 10–18, 1983.

[26] B. Tóth and G. Németh, "Improvements of Hungarian hidden Markov model-based text-to-speech synthesis," *Acta Cybern.*, vol. 19, no. 4, pp. 715–731, 2010.

[27] M. D. Zeiler, "ADADELTA: An Adaptive learning rate method," in *arXiv preprint*, 2012, p. arXiv:1212.5701.

[28] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011, vol. 15, pp. 315–323.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *arXiv preprint*, 2015, p. arXiv:1502.01852.

[30] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 249–256.

[31] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*, G. B. Orr and K.-R. Müller, Eds. Springer Berlin Heidelberg, 1998, pp. 9–50.

[32] G. Olaszy, "Precíziós, párhuzamos magyar beszédadatbázis fejlesztése és szolgáltatásai [Development and services of a Hungarian precisely labeled and segmented, parallel speech database] (in Hungarian)," *Beszédkutatás 2013 [Speech Res. 2013]*, pp. 261–270, 2013.

[33] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A Matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, 2011.

[34] D. J. Hermes, "Measuring the perceptual similarity of pitch contours," *J. Speech, Lang. Hear. Res.*, vol. 41, no. 1, pp. 73–82, Feb. 1998.

[35] "ITU-R Recommendation BS.1534: Method for the subjective assessment of intermediate audio quality." 2001.

[36] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.