PREDICTION OF VOICING AND THE F0 CONTOUR FROM ELECTROMAGNETIC ARTICULOGRAPHY DATA FOR ARTICULATION-TO-SPEECH SYNTHESIS

Simon Stone, Philipp Schmidt, Peter Birkholz

Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany

ABSTRACT

Articulation-to-speech synthesis based solely on supraglottal articulation requires some sort of intonation control. This paper examines to what extent the f_0 contour of an utterance can be predicted from such supraglottal articulation data. To that end, three groups of machine learning models (support vector machines, kernel ridge regression and neural networks) were trained and evaluated on the `mngu0` speech corpus containing synchronous articulatory and audio data. The best voiced/unvoiced/silence classification rates were achieved by a deep neural network with two hidden layers: 85.8 % with no look-ahead (important for on-line applications) and 86 % with a look-ahead of 50 ms. The best f_0 prediction model without look-ahead scored a root-mean-square error (RMSE) (when compared to the original f_0 contours) of 10.4 Hz using a neural network with one hidden layer, while the best prediction with a look-ahead of 50 ms was attained by kernel ridge regression and an RMSE of 10.3 Hz. The predicted f_0 contours were also subjectively evaluated in a listening test by manipulating the f_0 of the original speech files using PRAAT. The results are consistent with the objective evaluation.

***Index Terms*—** articulation-to-speech synthesis, intonation modeling, silent speech

1. INTRODUCTION

Articulation-to-speech synthesis is the process of generating artificial speech based on articulatory movements on-line and in real time: Some kind of articulometric technique (e.g., electromyography [1] or permanent magnetic articulography [2]) is used to record the speech *movements* (no sound production necessary) and the articulatory data is used to drive some kind of speech synthesizer (e.g., a vocoder [2, 3]) that outputs the intended speech. While it seems intuitively possible to obtain the speech sounds from their articulation, other speech characteristics like intonation or, more specifically, the fundamental frequency f_0, are inherent to the voice *source*, about which no information is known in this setup. However, recent studies have shown that the f_0 contour can nevertheless be derived from the articulatory data directly (e.g., [1, 4, 5, 6, 7]), from respiration [8], or by predicting the parameters of an intonation model [9]. While those studies showed promising results, they used different data sets, classifiers, regression models, different strategies for dealing with unvoiced segments and non-speech segments, and performed several predictions at the same time (e.g., a full articulation-to-speech mapping), which makes a reproduction and comparison of their results difficult. Given the focus on deep neural networks in most current studies, it was also of interest if a simpler model architecture that is less complex to train and has fewer hyperparameters to tune may be a good alternative in case of limited resources. This paper therefore systematically explores the performance of a set of commonly used models on the freely available `mngu0` corpus [10] containing synchronous speech audio and electromagnetic articulography data. Both the ternary classification of silent, voiced and unvoiced segments as well as the prediction of the f_0 contour were performed and evaluated.

2. METHODS AND MODELS

In a setup described above, where only the supraglottal articulation is measured and used to drive a speech synthesizer, no information on the voicing or even the voice activity is available. This poses two problems: when should the synthesizer be started and stopped and when should the synthesizer produce the voiced or the unvoiced instance of the same supraglottal articulation (e.g., when should it output a /g/ vs /k/)?

While speech activity and voicing is a ternary classification problem, the prediction of the f_0 at any given point in time is a regression problem, since the pitch can take on any value in a continuous range between certain physiological constraints. Therefore, the f_0 contour was predicted using the same methods as for the voiced/unvoiced/silence classification but modified to be used for the regression problem.

The machine learning techniques applied in this study were all implemented using the C++ machine learning toolkit "dlib" [11]. While the toolkit offers a plethora of tools and algorithms, three of the most commonly used families of algorithms were chosen for this investigation: support vector

This work was funded by the German Federal Ministry of Education and Research, reference number 13GW0101B, and the German Federal Ministry for Economic Affairs and Energy, reference number ZF4443004BZ8. The training and testing of all models were performed on a Bull Cluster at the Center for Information Services and High Performance Computing (ZIH) at TU Dresden.

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

machines (SVM), kernel ridge regression (KRR), and (deep) neural networks (DNN). For both investigated problems, the kernel functions used for the SVM and KRR were a linear kernel (i.e., no non-linear projection, LK), the radial basis function kernel (RBK), and the histogram intersection kernel (HIK). The DNNs were trained with one, three and five hidden layers with 512 neurons in each layer (to study the impact of the depth), and with two hidden layers with 512 neurons in the first and 1024 in the second layer (to mirror the configuration used by [5]), all of them using a Rectified Linear Unit (ReLU) activation function [12]. In addition to the properties described above, the hidden layers of these DNNs were fully connected, i.e., every neuron in each layer was connected to every other neuron in the next layer.

3. TRAINING

To train and evaluate the machine learning techniques described above, a corpus of articulatory data with corresponding pitch contours is required, since all of these techniques are supervised methods. One such corpus is the publicly available `mngu0` corpus [10] containing synchronous speech audio and, among other forms of articulatory data, electromagnetic articulography data of one speaker from two recording sessions.

3.1. Dataset

We used the Day 1 set of EMA data along with the corresponding audio data and extended SAMPA annotation for training and evaluation of the classification and regression models. The Day 1 set contains 1 354 utterances by one male British professional speaker amounting to 67 min of speech data, which were randomly split (without breaking up utterances) into a training set (80 % of the total data frames) and a test set (the remaining 20 % of the data). The sentence lengths ranged from 1 to 48 words and included questions, statements and exclamations (and therefore a variety of intonation contours). In total, the set contained approximately 1 715 unique diphones and 12 322 unique triphones. The EMA data was sampled using the Carstens AG500 articulograph, which is capable of tracking 12 EMA sensor coils in 3D space with two angles of rotation for a total of 5 measurements per sensor coil. For this study, we used only the x- and y-coordinates of six coils (placed on the upper lip, the lower lip, the lower incisor, the tongue tip, the tongue body, and the tongue dorsum) in the midsagittal plane for a total of 2 x 6 = 12 channels (number of dimensions times number of coils). This limitation was imposed to remain within the subset of data that the authors of the corpus have already evaluated and processed themselves: The `mngu0` corpus contains processed EMA data of these 12 channels. Due to possible overlap of some of the corresponding audio files (according to the dataset's readme file), we used the raw data and performed our own processing by normalizing the EMA data channel-wise so that each channel (containing data representing one spatial dimension of one coil) exhibited a mean of 0 and a standard deviation of 1. The corpus contains two sets of audio recordings: one recorded using a Sennheiser MKH50 hypercardioid, which picked up background noise from the AG500 starting at about 7.5 kHz, and a PHON-OR noise-cancelling optical microphone, which had a smaller bandwidth and did not pick up low frequencies very well. Because this study was interested in the fundamental frequency, the Sennheiser MKH50 audio recordings were used for training and evaluating the f_0 prediction as the noise interference was well above the expected frequency range of interest.

3.2. Articulatory feature vectors

The data were presented to the machine learning algorithms as a series of feature vectors, each of which represented one frame of EMA data sampled every 5 ms. The feature vectors consisted of the 12 channel data in that frame (x- and y- coordinates of the six sensor coils as described above), the element-wise, normalized difference of the current 12 channel data to the previous 12 channel data (i.e., delta features), and the element-wise, normalized difference of the current difference to the previous difference (i.e., delta-delta features). In total, each feature vector therefore had a length of 36. To include an articulatory context for each feature vector, several consecutive feature vectors were stacked. Two different kinds of context were studied: using only previous feature vectors and additionally using subsequent feature vectors, as well. The former case is feasible in a real-time articulation-to-speech synthesis system as described above, while the latter setting was expected to improve the results at the cost of a small delay. Context lengths of 25 ms, 50 ms, and 75 ms were studied, but for the sake of conciseness, we only report the best results here, which were achieved with a context of 50 ms corresponding to 10 frames for both look-back and look-ahead.

3.3. Extraction of the reference f_0 contour

To train the supervised machine learning models used in this study, each training frame was assigned a label for unvoiced/voiced/silence classification and an f_0 value for regression. The silence label can be directly extracted from the extended SAMPA annotation of the `mngu0` corpus. But since no narrow transcription of the utterances was available, we based the voiced/unvoiced label on the results of the f_0 extraction: if no f_0 could be determined, the frame was labeled "unvoiced", otherwise it was considered "voiced". To determine the f_0 of each non-silent frame, we used PRAAT's [13] autocorrelation-based *To Pitch...* function with a pitch floor of 50 Hz and a pitch ceiling of 200 Hz. If the PRAAT algorithm could not find a sufficiently clear peak in the autocorrelation-function, it returned the value "undefined" for that frame. This was replaced by the numeric value -1 and used as the unvoiced-flag so that for the voiced-unvoiced classification all positive values were interpreted as a "voiced" label. In total, 363 502 voiced samples and 368 021 unvoiced samples were used for training, and 126 651 voiced frames and 121 965 unvoiced and silent frames for testing the classifiers. The regression models were trained and tested with the voiced frames only, although unvoiced or silent frames were included in the articulatory feature vectors if they appeared in the context of a voiced frame.

3.4. Data partitioning strategy and hyperparameter tuning

Even though the mngu0 corpus proposes a standard split for training, validation and testing, we chose to make our own split, following a practice suggested by [14]. As described above, 80 % of the feature vectors and their corresponding labels were used for the training of all investigated classifiers and regression models and 20 % of the data was kept separately for testing. The validation set for hyperparameter tuning (as a subset of the training set) was determined differently for each class of models: The neural networks were trained using a mini-batch paradigm with a batch size of 512. The learning rate was the only hyperparameter that was tuned. Its optimum was found by successively shrinking the learning rate from 0.1 by a factor of 10 after every training epoch and evaluation on the test set. The hyperparameters of the SVMs were optimized using a grid search of the parameter space and a two-fold cross-validation on the training set. Due to the large training set, a higher number of folds was too computationally expensive. The hyperparameter λ for the KRR was found using leave-one-out cross-validation since training was much faster and thus allowed a more thorough cross-validation. The final hyperparameter values for all models are summarized in Table 1 and Table 2.

4. RESULTS

After training the models on the training sets using the optimal hyperparameters shown in Tables 1 and 2, the trained models were then evaluated on the respective test sets. The evaluation was performed using both objective measures (the classification score and the regression error) and a subjective listening test using human listeners.

4.1. Objective evaluation

The classification score was calculated by dividing the number of correctly classified voiced and unvoiced frames by the total number of frames in the test set. The regression error was determined in terms of the root mean square error (RMSE) between the predicted f_0 and the reference f_0 determined with PRAAT (see above). The results of the evaluation are shown in Figures 1 and 2. It is evident that in both settings the non-linear kernel methods outperformed the linear methods. The DNNs also generally slightly outperformed SVM and KRR models for classification. For the regression task, the DNNs are generally on par with the SVR or KRR models with a KRR using a radial basis kernel achieving the overall best result of an RMSE of 10.3 Hz. This is somewhat surprising, given the dominance of deep networks in almost every field. An explanation could be the vast number of possible network topologies and hyperparameter settings of a DNN. Even with the careful approach taken here, there is no guarantee that the true global optimum was found. Compared to the previous benchmark set by [5], which was an RMSE of 12.6 Hz achieved on the same corpus using a long short-term memory (LSTM), the results from our study are an improvement of approximately 17 %. However, another study using an LSTM [7] performed slightly better with a reported RMSE of 10.162 Hz when using the same input data as in our study, most likely due to using a (well-tuned) LSTM instead of a simple feed-forward DNN.

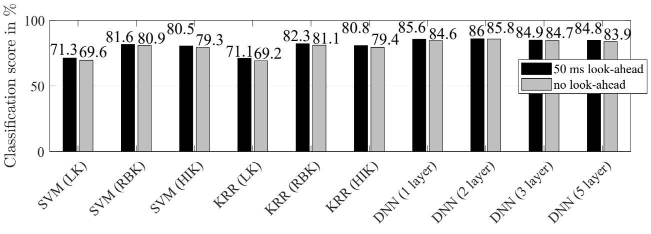

Fig. 1. Voiced/unvoiced/silence classification score in percent of correctly classified frames.

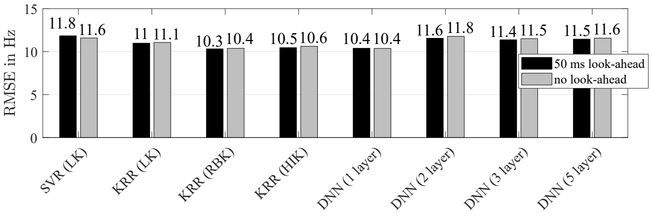

Fig. 2. Root mean square error (RMSE) of predicted f_0 contour with respect to the reference f_0.

While adding a look-ahead for the articulatory context improved the results marginally, the proposed techniques are still sufficiently precise for a real-time application in an articulation-to-speech synthesis system even when no "future" context is used.

4.2. Listening test

While the RMSE is a commonly used objective measure to evaluate a regression model and a good index to compare different algorithms, it is not intuitively clear how it relates to perceived quality or naturalness of the produced contours. We therefore conducted a listening test where a subset of the results was rated by 20 human listeners (8 female, 12 male, age 22-56, average age 30.4). To limit the number of the stimuli

	SVM				KRR				DNN
	LK	RBK		HIK	LK	RBK		HIK	all
Context (before | after) in ms	C	C	γ	C	λ	λ	γ	λ	LR
50 | 0	2e+06	9.842e+05	0.0045	3e+05	0.0001	0.01	0.0081	0.01	0.0001
50 | 50	2e+06	2e+06	0.0015	3e+05	0.0001	0.001	0.0009	1	0.0001

Table 1. Optimal hyperparameters for the silence/voiced/unvoiced classifiers. The optimal learning rate (LR) was the same for every number of hidden layers of the DNN.

	SVR	KRR				DNN
	LK	LK	RBK		HIK	all
Context (before | after) in ms	C	λ	λ	γ	λ	LR
50 | 0	1.968e+05	1e-05	0.01	0.0009	0.01	1e-07
50 | 50	5e+05	1e-05	0.1	0.0009	0.01	1e-07

Table 2. Optimal hyperparameters for the f_0 regression. The optimal learning rate (LR) was the same for every number of hidden layers of the DNN.

to a reasonable amount, we grouped the utterances from the test set into tertiles using the RMSE: the best (lowest RMSE) third (T1), the median third (T2), and worst (highest RMSE) third (T3). We then randomly selected one short, one long and one medium long utterance from each third. For each of these nine utterances, we manipulated the f_0 contour in the original audio recordings according to the f_0 predicted by the best regression model with look-ahead and without look-ahead using PRAAT. We also added a sample of each utterance with a completely flattened intonation (setting it to its mean f_0) and the unmodified recordings of each sample[1]. The resulting 9 utterances × 4 versions = 36 utterances were presented to the subjects three times in a randomized fashion for a total of 108 items per test. The items were presented to the listeners in a quiet room using a Focusrite Saffire Pro 40 audio interface and a pair of Beyerdynamic T70p headphones. The raters were asked to grade each item on a scale from 1 (unnatural) to 4 (very natural). The results of the test are shown in Figure 3. The original f_0 contours and the flattened f_0 contours scored highest and lowest, as could be expected. The ratings of the predicted f_0 contours were also consistent with the objective evaluation. The selected sentences are available as supplemental material accompanying this article and at http://www.vocaltractlab.de/index.php?page=birkholz-supplements.

5. SUMMARY AND CONCLUSION

We conducted a systematic comparison of a number of machine learning algorithms' (SVM, KRR, and DNN) performances for f_0 prediction from articulatory data. Our results show that DNNs are generally a good option for both classification of voiced/unvoiced frames and predicting the f_0 in voiced frames. The results were only marginally worse when using only the current feature vector and the previous 50 ms of data as opposed to a look-ahead of 50 ms, indicating the suitability of the proposed methods for a real-time system. The best classification score was 86 % and achieved by a DNN with two hidden layers and 50 ms of context both before and after the frame of interest. The lowest prediction error was 10.3 Hz and achieved using symmetric context of 50 ms and KRR with a radial-basis function kernel.

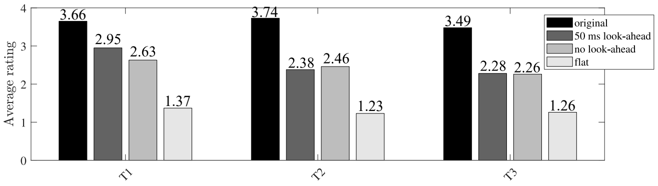

Fig. 3. Average naturalness rating in the listening test of the resynthesized utterances using the orignal, predicted, and flattened f_0 contours.

A significant finding of this study was that even drastically simpler techniques like KRR can achieve performance meeting or exceeding the performance of a DNN, which needs large amounts of data, demands computationally expensive training and is notoriously difficult to optimize. Another significant finding was the fact that not using a look-ahead did not significantly decrease the performance of both unvoiced/voiced/silence classification and f_0 regression. This is an important fact for the design of real-time articulation-to-synthesis systems.

The results from the listening test suggest that the RMSE is a valid error measure since the tested items were rated in the same relative ranked order than their respective RMSE. But even though listeners judged the predicted f_0 contours as acceptably natural, it remains to be investigated if pitch accents could also be learned from the supraglottal articulation and then correctly synthesized. Similarly, more expressive intonation should be analyzed, as the dataset underlying this study only included read speech with neutral emotion.

[1] "Unmodified" means that we passed it through PRAAT's pitch manipulation algorithm once without changing anything. This seemingly redundant step was necessary because the pitch manipulation algorithm introduces a small amount of noise that would otherwise unfairly skew the comparison.

6. REFERENCES

[1] Keigo Nakamura, Matthias Janke, Michael Wand, and Tanja Schultz, "Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, Prague, Czech Republic, 2011, pp. 573–576.

[2] Jose A Gonzalez, Lam A Cheah, Angel M Gomez, Phil D Green, James M Gilbert, Stephen R Ell, Roger K Moore, and Ed Holdsworth, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2362–2374, 2017.

[3] Lorenz Diener and Tanja Schultz, "Investigating objective intelligibility in real-time EMG-to-Speech conversion," in *Proc. of the Interspeech*, Hyderabad, India, 2018, pp. 3162–3166.

[4] Zheng-Chen Liu, Zhen-Hua Ling, and Li-Rong Dai, "Articulatory-to-acoustic conversion with cascaded prediction of spectral and excitation features using neural networks," in *Proc. of the Interspeech*, San Francisco, USA, 2016, pp. 1502–1506.

[5] Cenxi Zhao, Longbiao Wang, Jianwu Dang, and Ruiguo Yu, "Prediction of F0 based on articulatory features using DNN," in *Proc. of the International Seminar on Speech Production (ISSP)*, Tianjin, China, 2017, pp. 58–67.

[6] Tamás Grósz, Gábor Gosztolya, László Tóth, Tamás Gábor Csapó, and Alexandra Markó, "F0 estimation for DNN-based ultrasound silent speech interfaces," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 291–295.

[7] Beiming Cao, Myungjong Kim, Jun R Wang, Jan van Santen, Ted Mau, and Jun Wang, "Articulation-to-speech synthesis using articulatory flesh point sensors' orientation information," in *Proc. of the Interspeech*, Hyderabad, India, 2018, pp. 3152–3156.

[8] Farzaneh Ahmadi and Tomoki Toda, "Designing a pneumatic bionic voice prosthesis - a statistical approach for source excitation generation," in *Proc. of the Interspeech*, 2018, pp. 3142–3146.

[9] Bastian Schnell and Philip N. Garner, "A neural model to predict parameters for a generalized command response model of intonation," in *Proc. of the Interspeech*, Hyderabad, India, 2018, pp. 3147–3151.

[10] Korin Richmond, Phil Hoole, and Simon King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Proc. of the Interspeech*, Florence, Italy, 2011, pp. 1505–1508.

[11] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[12] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve Restricted Boltzmann Machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[13] Paul Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, pp. 341–345, 2002.

[14] Kyle Gorman and Steven Bedrick, "We need to talk about standard splits," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019, pp. 2786–2791, Association for Computational Linguistics.