A new Pitch Tracking Smoother based on Deep Neural Networks

Michele Ferro
FICLIT, University of Bologna, Italy
lele.ferro4@gmail.com

Fabio Tamburini
FICLIT, University of Bologna, Italy
fabio.tamburini@unibo.it

Abstract

English. This paper presents a new pitch tracking smoother based on deep neural networks (DNN). The proposed system has been extensively tested using two reference benchmarks for English and exhibited very good performances in correcting pitch detection algorithms outputs.

Italiano. *Questo contributo presenta un programma di* smoothing *del profilo intonativo basato su reti neurali* deep. *Il sistema è stato verificato utilizzando due corpora di riferimento e le sue prestazioni nella correzione degli errori di alcuni algoritmi per l'identificazione del pitch sono decisamente buone.*

1 Introduction

The pitch, and in particular the fundamental frequency - F0 - which represents its physical counterpart, is one of the most relevant perceptual parameters of the spoken language and one of the fundamental phenomena to be carefully considered when analysing linguistic data at a phonetic and phonological level. As a consequence, the automatic extraction of F0 has been a subject of study for a long time and in literature there are many works that aim to develop algorithms able to reliably extract F0 from the acoustic component of the utterances, algorithms that are commonly identified as Pitch Detection Algorithms (PDAs).

Technically, the extraction of F0 is a problem far from trivial and the great variety of methodologies applied to this problem demonstrates its extreme complexity, especially considering that it is difficult to design a PDA that works optimally for the different recording conditions, considering that parameters such as speech type, noise, overlap, etc. are able to heavily influence the performance of this type of algorithms.

Scholars worked hard searching for increasingly sophisticated techniques for these particular cases, although extremely relevant for the construction of real applications, considering solved, or perhaps simply abandoning, the problem of the F0 extraction for the so-called "clean speech". However, anyone who has used the most common programs available for the automatic extraction of F0 is well aware that errors of halving or doubling of the value of F0, to cite only one type of problem, are far from rare and that the automatic identification of voiced areas within the utterance still poses numerous problems.

Every work that proposes a new method for the automatic extraction of F0 should perform an evaluation of the performances obtained in relation to other PDAs, but, usually, these assessments suffer from the typical shortcomings deriving from evaluation systems: they usually examine a very limited set of algorithms, often not available in their implementation, typically considering corpora not distributed, related to specific languages and/or that contain particular typologies of spoken language (pathological, disturbed by noise, etc.) (Veprek, Scordilis, 2002; Wu *et al.*, 2003; Kotnik *et al.*, 2006; Jang *et al.*, 2007; Luengo *et al.*, 2007; Chu, Alwan, 2009; Bartosek, 2010; Huang, Lee, 2012; Chu, Alwan, 2012). There are few studies, among the most recent, that have performed quite complete evaluations that are based on corpora freely downloadable (deCheveigné, Kawahara, 2002; Camacho, 2007; Wang, Loizou, 2012). These studies use very often a single metric in the assessment that measures a single type of error, not considering or partly considering the whole panorama of indicators developed from the pioneering work of Rabiner and colleagues (1976) and therefore, in our opinion, the results obtained seem to be rather partial.

Tamburini (2013) performed an in depth study of the different performances exhibited by several

widely used PDAs by using standard evaluation metrics and well established corpus benchmarks.

Starting from this study, the main purpose of our research was to improve the performances of the best Pitch Detection Algorithms identified in Tamburini (2013) by introducing a post-processing smoother. In particular, we implemented a pitch smoother adopting Keras[1], a powerful high-level neural networks application program interface (API), written in Python and capable of running on top of TensorFlow, CNTK, or Theano.

2 Pitch error correction and smoothing

Typical PDAs are organised into two different modules: the first stage tries to detect pitch frequencies frame by frame and, in the second stage, the pitch candidates or probabilities are connected into pitch contours using dynamic programming techniques (Bagshaw, 1994; Chu, Alwan, 2012; Gonzalez, Brookes, 2014) or hidden Markov models (HMMs) (Jin, Wang, 2011; Wu *et al.*, 2003).

These techniques are, however, not completely satisfactory and various kind of errors remain in the intonation profile. That is why in the literature we can find various studies aiming at proposing pitch profile smoothers. Some works try to correct intonation profile by applying traditional techniques (Zhao *et al.*, 2007; So *et al.*, 2017; Jlassi *et al.*, 2016), while few others (see for example (Kellman, Morgan, 2016; Han, Wang, 2014)) are based on DNN (either Mulity-Layer Perceptrons or Elman Recurrent Neural Networks).

The pitch smoother we propose is based on recurrent neural networks in order to process the entire sequence of raw pitch values computed by the various PDAs and trying to correct it by removing, mainly, halving/doubling errors and other kind of glitches that could appear in raw pitch profiles.

At the input layer we inject one-hot vectors representing the frame pitch value in the interval 0-499Hz as detected by the PDA. We kept the pitch frame size required by each PDA imposing only a frame shift of 0.01 sec for every PDA. With regard to the hidden layer we employed a bidirectional Long-Short-Term Memory (LSTM) with 100 neurons for each direction. They are joined together and inserted into a TimeDistributed wrapper layer so that one value per timestep could be predicted (instead getting one value for each sequence) given the full sequence of one-hot vectors provided as input.

At the output softmax layer we expect to get a probability distribution for the pitch values in the same interval 0-499Hz, considering the most likely one as the actual network prediction. This means that the network input and output layers contain 500 neurons each.

3 Experiments setup

3.1 Tested PDAs

We chose the three PDAs exhibiting the best performances in Tamburini (2013), namely RAPT, SWIPE' and YAAPT. Even though they were originally developed as MATLAB functions, we decided to adopt the corresponding Python implementations.

The primary purpose in the development of RAPT (A Robust Algorithm for Pitch Tracking) (Talkin, 1995) was to obtain the most robust and accurate estimates possible, with little thought to computational complexity, memory requirements or inherent processing delay. This PDA is designed to work at any sampling frequency and frame rate over a wide range of possible F0, speaker and noise condition. For the determination of the pitch profile, a Normalized Cross-Correlation Function (NCCF) is used and each candidate of F0 is estimated thanks to dynamic programming techniques. The Python implementation is available at `http://sp-tk.sourceforge.net/`.

SWIPE (The Sawtooth Inspired Pitch Estimator) (Camacho, 2007) improves the performance of pitch tracking adopting these measures: it avoids the use of the logarithm of the spectrum, it applies a monotonically decaying weight to the harmonics, then the spectrum in the neighbourhood of the harmonics and middle points between harmonics are observed and smooth weighting functions are used. We adopted SWIPE', a variant of this PDA that only uses the main harmonics for pitch estimation, implemented in Python and it is available again at `http://sp-tk.sourceforge.net/`.

The YAAPT (Yet Another Algorithm for Pitch Tracking) (Zahorian, Hu, 2007) is a fundamental frequency (Pitch) tracking algorithm, which is designed to be highly accurate and very robust for both high quality and telephone speech. In gen-

[1] https://keras.io/

eral, a preprocessing step is used to create multiple versions of the signal. Consequently, spectral harmonics correlation techniques (SHC) and a Normalized Cross-Correlation Function (NCCF, as in RAPT) are adopted. The final profile of F0 is estimated thanks to dynamic programming techniques. For our experiments we employed pYAAPT, a Python implementation available at `http://bjbschmitt.github.io/AMFM_decompy/pYAAPT.html`.

3.2 Gold Standards

The evaluation tests were based on two English corpora considered as gold standards, both freely available and widely used in literature for the evaluation of PDAs:

- Keele Pitch Database (Plante *et al.*, 1995): it is composed of 10 speakers, 5 males and 5 females, who read, in a controlled environment, a small balanced passage (the 'North Wind story'). The corpus contains also the output of a laryngograph, from which it is possible to accurately estimate the value of F0.

- FDA (Bagshaw *et al.*, 1993): it is a small corpus containing 5' of recording divided into 100 utterances, read by two speakers, a male and a female, particularly rich in fricative sound, nasal, liquid and glide, sounds particularly problematic to be analysed by the PDAs. Also in this case the gold standard for the values of F0 is estimated starting from the output of the laryngograph.

3.3 Evaluation metrics

Proper evaluation mechanisms have to introduce suitable quantitative measures of performance that should be able to grasp the different critical aspects of the problem under examination. In Rabiner *et al.* (1976) a de facto standard for PDA assessment measures is established, a standard used by many others after him (e.g. (Chu, Alwan, 2009)). If $E_{voi \to unv}$ and $E_{unv \to voi}$ respectively represent the number of frames erroneously classified between voiced and unvoiced and vice versa, while E_{f0} represents the number of voiced frames in which the pitch value produced by the PDA differs from the gold standard for more than 16Hz, then we can define:

- Gross Pitch Error:

$$GPE = E_{f0}/N_{voi}$$

- Voiced Detection Error:

$$VDE = (E_{voi \to unv} + E_{unv \to voi})/N_{frame}$$

where N_{voi} is the number of voiced frames in the gold standard and N_{frame} is the number of frames in the utterance. These indicators, taken individually or in pairs, have been used in a large number of works to evaluate the performance of PDAs. The two indicators, however, measure very different errors; it is possible to measure the performance using only one indicator, usually GPE, but it evaluates only part of the problem and hardly provide a faithful picture of PDA behaviour. On the other hand, considering both measures leads to a difficult comparison of the results.

To try to remedy these problems, Lee and Ellis (2012) have suggested slightly different metrics, which allow the definition of a single indicator:

- Voiced Error:

$$VE = (E_{f0} + E_{voi \to unv})/N_{voi}$$

- Unvoiced Error:

$$UE = E_{unv \to voi}/N_{unv}$$

- Pitch Tracking Error:

$$PTE = (VE + UE)/2$$

where N_{unv} is the number of unvoiced frames contained in the gold standard. However, trying to interpret the results obtained by a PDA in light of the PTE measurement is rather complex: it is not immediate to identify from the obtained results the most relevant source of errors.

In the light of what has been said so far, it seems appropriate to introduce a new measure of performance that is able to easily capture the performance of a PDA in a single, clear indicator that considers all types of possible errors to be equally relevant. So, following Tamburini (2013), we adopt, the Pitch Error Rate as performance metric, defined as:

$$PER = (E_{f0} + E_{voi \to unv} + E_{unv \to voi})/N_{frame}$$

This measure sum all the types of possible errors without privileging or reducing the contribution of any component and allowing a simpler interpretation of the obtained outcomes.

4 Results

We repeated the same experiments as in Tamburini (2013) with the Python implementations of the chosen algorithms (See Table 1) in order to derive common baselines. We also computed the median of the values as in Tamburini (2013) as a simple smoothing method. As in the cited work, it emerges quite clearly that the combination of different algorithms with the median method improves the PER results.

Keele Pitch Database				
PDA	PER	E_{f0}	$E_{voi \to unv}$	$E_{unv \to voi}$
pYAAPT	0.14056	0.04278	0.04411	0.05366
RAPT	0.12596	0.03789	0.05252	**0.03554**
SWIPE'	0.14236	0.02762	0.06985	0.04488
Median	**0.08814**	**0.02656**	**0.03359**	0.03564
FDA Corpus				
PDA	PER	E_{f0}	$E_{voi \to unv}$	$E_{unv \to voi}$
pYAAPT	0.11912	0.03023	**0.03399**	0.0549
RAPT	0.09533	0.01978	0.03438	0.04116
SWIPE'	0.10594	**0.01385**	0.04773	0.04434
Median	**0.10182**	0.02537	0.03686	**0.03917**

Table 1: The experiments in Tamburini (2013) reproduced using the considered PDA python implementation.

After the influential paper from Reimers and Gurevych (2017) it is clear to the community that reporting a single score for each DNN training session could be heavily affected by the system initialisation point and we should instead report the mean and standard deviation of various runs with the same setting in order to get a more accurate picture of the real systems performances and make more reliable comparisons between them.

In order to carry out the experiments with our new pitch smoother we had to split our datasets into training/validation/test set. For the final evaluation of our pitch smoother, we considered only the PER measure. This metric was computed for each epoch during the training phase for all subsets in order to determine the stopping epoch when we get the minimum PER on the validation set. We performed 10 runs for each experiment computing means, standard deviations and significance tests.

We also tested our pitch smoother on a mixed configuration joining our datasets and adopting the same procedures.

Table 2 shows all the obtained results. The proposed system always exhibits the best results in any experiment with relevant performance gains with respect to the PDAs base outputs. All the differences resulted highly significant when applying a t-test. Given the very small standard deviation in all the experiments we can conclude that, in this case, the initialisation point did not affect the network performances too much.

Keele Pitch Database			
PDA	PDA PER	Smoother PER μ	Smoother PER σ
pYAAPT	0.14056	**0.05458**	0.00157
RAPT	0.12596	**0.08726**	0.00193
SWIPE'	0.14236	**0.09666**	0.00298
FDA Corpus			
PDA	PDA PER	Smoother PER μ	Smoother PER σ
pYAAPT	0.11912	**0.06530**	0.00277
RAPT	0.09533	**0.06698**	0.00133
SWIPE'	0.10594	**0.07205**	0.00215
Mixed Keele+FDA Corpus			
PDA	PDA PER	Smoother PER μ	Smoother PER σ
pYAAPT	0.06951	**0.05415**	0.00128
RAPT	0.09859	**0.07341**	0.00133
SWIPE'	0.08758	**0.08288**	0.00163

Table 2: PER mean (μ) and standard deviation (σ) obtained by the proposed pitch profile smoother. One sample t-test significance test returns $p \ll 0.001$ for all experiments. N.B.: Even if the number of experiments is small (10), the power analysis of the t-tests is always equal to 1.0 showing maximum t-test reliability.

5 Conclusions

This paper presented a new pitch smoother based on deep neural networks that obtained excellent results when evaluated using standard benchmarks for English and evaluation metrics proposed in the literature.

Future works could regard the intermixing of various corpora in different languages in order to test the possibility of deriving a pitch smoother able to properly work without caring about language and, possibly, specific corpora and language registers.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Ti-

tan Xp GPU used for this research.

References

Bartosek, J. 2010 Pitch Detection Algorithm Evaluation Framework *In Proceedings of 20th Czech-German Workshop on Speech Processing*, Prague, 118123.

Bagshaw, P.C. 1994 *Automatic prosodic analysis for computer-aided pronunciation teaching*, PhD Thesis, University of Edimburgh.

Bagshaw, P.C. and Hiller, S.M. and Jack, M.A. 1993 Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching, *Proceedings of Eurospeech '93*, Berlin, 1003–1006

Camacho A. 2007 SWIPE: A sawtooth waveform inspired pitch estimator for speech and music. *PhD Thesis, University of Florida*.

Chu, W. and Alwan A. 2009 Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend *In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP2009*, 39693972.

Chu, W. and Alwan, A. 2012. SAFE: A statistical approach to F0 estimation under clean and noisy conditions. *IEEE Trans. Audio, Speech, Lang. Process.*, 20(3):933–944.

de Cheveigné A. and Kawahara H. 2002 YIN, a fundamental frequency estimator for speech and music *Journal of the Acoustical Society of America*, 111, 191730.

Gonzalez, S. and Brookes, M. 2014. PEFAC-A pitch estimation algorithm robust to high levels of noise. *IEEE Trans. Audio, Speech, Lang. Process.*, 22(2):518–530.

Han, Kun and Wang, DeLiang 2014. Neural Network Based Pitch Tracking in Very Noisy Speech. *IEEE Trans. Audio, Speech, Lang. Process.*, 22(12):2158–2168.

Huang, F. and Lee, T. 2012 Robust Pitch Estimation Using l1-regularized Maximum Likelihood Estimation. *In Proceedings of 13th Annual Conference of the International Speech Communication Association Interspeech 2012*, Portland (OR).

Jang, S.J. and Choi, S.H. and Kim, H.M. and Choi, H.S. and Yoon Y.R. 2007 Evaluation of performance of several established pitch detection algorithms in pathological voices. *In Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society - EMBC*, Lyon, 620623.

Jin, Z. and Wang, L. 2011. HMM-based multipitch tracking for noisy and reverberant speech. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(5):1091–1102.

Jlassi, Wided and Bouzid, Aicha and Ellouze, Noureddine 2016 A new method for pitch smoothing, *2nd International Conference on Advanced Technologies for Signal and Image Processing*, Monastir, Tunisia, 657–661.

Kellman. M. and Morgan, N. 2017 Robust Multi-Pitch Tracking: a trained classifier based approach, *ICSI Tchnical Report*, Berkeley, CA.

Kotnik, B. and Höge, H. and Kacic, Z. 2006 Evaluation of Pitch Detection Algorithms in Adverse Conditions *In Proceedings of Speech Prosody 2006*, Dresden, PS2883.

Lee, B.S. and Ellis, D. 2012 Noise Robust Pitch Tracking by Subband Autocorrelation Classification *In Proceedings of 13th Annual Conference of the International Speech Communication Association Interspeech 2012*, Portland (OR).

Luengo, I., Saratxaga, I., Navas, E., 2007 Evaluation of Pitch Detection Algorithm under Real Conditions. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*, Honolulu, Hawaii, 4, 10571060.

Plante, F. and Ainsworth, W.A. and Meyer, G. 1995 A Pitch Extraction Reference Database. *In Proceedings of Eurospeech95*, Madrid, 837840.

Rabiner, L.R. and Cheng, M.J. and Rosenberg, A.E. and McGonegal C.A. 1976 A Comparative Performance Study of Several Pitch Detection Algorithms. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 24, 399418.

Reimers, Nils and Gurevych, Iryna. 2017 Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, 338–348.

So, YongJin and Jia, Jia and Cai, LianHong. 2012 Analysis and Improvement of Auto-correlation Pitch Extraction Algorithm Based on Candidate Set, In Zhihong Q., Lei C., Weilian S., Tingkai W., Huamin Y. (eds) *Recent Advances in Computer Science and Information Engineering: Volume 5*, Springer Berlin Heidelberg, 697–702.

Talkin D. 1995 A robust algorithm for pitch tracking (RAPT). In Kleijn W.B., Paliwal, K.K. (eds) *Speech Coding and Synthesis*, New York: Elsevier, 495518.

Tamburini, Fabio 2013 Una valutazione oggettiva dei metodi pi diffusi per l'estrazione automatica della frequenza fondamentale. *In Atti dell IX Convegno Nazionale dell'Associazione Italiana*

di Scienze della Voce (AISV2013), Bulzoni:Roma, 427–434.

Veprek, P. and Scordilis, M.S. 2002 Analysis, enhancement and evaluation of five pitch determination techniques. *Speech Communication*, 37, 249270.

Wang, D. and Loizou, P.C. 2012 Pitch Estimation Based on Long Frame Harmonic Model and Short Frame Average Correlation Coefficient. *In Proceedings of 13th Annual Conference of the International Speech Communication Association Interspeech 2012*, Portland (OR).

Wu, M. and Wang, L. and Brown G.J. 2003. A multipitch tracking algorithm for noisy speech. *IEEE Trans. Audio, Speech, Lang. Process.*, 11(3):229–241.

Zahorian, S.A. and Hu, H. 2008 A Spectral/temporal method for Robust Fundamental Frequency Tracking. *Journal of the Acoustical Society of America*, 123, 45594571.

Zhao, Xufang and O'Shaughnessy, Douglas and Minh-Quang, Nguyen. 2007 A Processing Method for Pitch Smoothing Based on Autocorrelation and Cepstral F0 Detection Approaches, *Proceedings of the International Symposium on Signals, Systems and Electronics*, Montreal, Canada, 59–62