Postprint of: M. Blok, J. Banaś and M. Pietrolaj, "IFE: NN-aided Instantaneous Pitch Estimation," 2021 14th International Conference on Human System Interaction (HSI), 2021, pp. 1-7, doi: 10.1109/HSI52170.2021.9538713

IFE: NN-aided Instantaneous Pitch Estimation

Marek Blok
Gdansk University of Technology
Faculty of Electronics,
Telecommunications and Informatics
Gdansk, Poland
Marek.Blok@pg.edu.pl

Jan Banaś
Gdansk University of Technology
Faculty of Electronics,
Telecommunications and Informatics
Gdansk, Poland
Jan.Banas@pg.edu.pl

Mariusz Pietrołaj
Gdansk University of Technology
Faculty of Electronics,
Telecommunications and Informatics
Gdansk, Poland
Mariusz.Pietrolaj@pg.edu.pl

2021 14th International Conference on Human System Interaction (HSI) | 978-1-6654-4112-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/HSI52170.2021.9538713

Abstract*— Pitch estimation is still an open issue in contemporary signal processing research. Nowadays, growing momentum of machine learning techniques application in the data-driven society allows for tackling this problem from a new perspective. This work leverages such an opportunity to propose a refined Instantaneous Frequency and power based pitch Estimator method called IFE. It incorporates deep neural network based pitch estimation with audio front end used for extraction of instantaneous frequency and power of signal components. A thorough results analysis is performed and major advantages and shortcomings of this method are identified, leading to a wide array of suggestions for future improvement. While IFE exhibits an instantaneous temporal resolution, a comparison is made against state-of-the-art pitch estimators operating on time windows, proving a comparable degree of prediction accuracy (up to 6% accuracy improvement) while maintaining the advantage of higher temporal resolution.

Keywords*— pitch estimation, machine learning, speech synthesis, data augmentation, neural network, IFE

I. Introduction

In terms of human-machine interaction, one of the preferred and rapidly disseminating interfaces is the speech interface – allowing for hands-free operation of systems and enabling human-like interaction. Accurate speech recognition and speaker identification require complex signal processing pipelines, including signal pre-conditioning, feature extraction, acoustic fingerprinting and others. One of the tools enabling these algorithms is pitch estimation – the process of identifying the fundamental frequency (F_0) of voiced speech. Many solutions to this problem have been proposed over the years, however few have approached the problem from instantaneous perspective. In this contribution, a novel solution is proposed - based on the authors' prior work, but offering significant refinement.

II. Analysis

A. State of the art

Pitch estimation has been one of the fundamental problems in sound analysis and information retrieval since the early days of sound recording and processing. Initial approaches to this task took advantage of the autocorrelation function (ACF) of a signal in time domain [1] as well as processing in frequency or cepstrum domains [2]. However, it wasn't until digital signal processing became ubiquitous that the results reached a usable level of performance in real-life scenarios [3]. Later on, with the rise of machine learning a new family of solutions emerged, eventually taking the form of a hybrid solution , combining best practices from the conventional digital signal processing (DSP) approaches with support from data-driven machine learning.

B. Conventional DSP

From among the impressive variety of conventional DSP pitch estimators, three relatively recent and popular algorithms were examined in the course of this effort.

A Robust Algorithm for Pitch Tracking (RAPT) [4] relies on computing a normalized cross correlation function (NCCF) for a significantly down-sampled version of the analyzed signal to identify a set of candidates for further search via NCCF of the original signal. A set of local maxima is selected as pitch candidates, from among which a dynamic programming solution is applied to pick the most likely one, based on local and contextual evidence.

The YIN pitch estimator [5] is an exceptionally accurate solution, especially given the fact it relies on the ACF of the signal with further time-domain processing and hence requires a considerably low amount of computational power. It was designed to handle both speech and music signals. Since its initial introduction, it has been also refined by other contributors, resulting in embodiments such as YIN-bird [6] – tuned for songbirds, or pYIN [7] – a refinement including temporal continuity enforcement by a hidden Markov model (HMM).

Pitch Estimation Filter with Amplitude Compression (PEFAC) [8] is the most recent of the presented conventional DSP pitch estimators. It operates in the frequency domain, where power spectral density (PSD) of the signal is convolved with a logarithmic comb-like analysis filter capable of pointing to the highest peak – thus identifying the most probable F_0 candidate. No temporal continuity enforcement is incorporated in this solution.

C. Machine learning

CREPE [9] is a well-established standalone machine learning solution to the pitch estimation problem. Within it, a convolutional neural network (CNN) is deployed with no DSP front-end or feature extraction – the input to the net is the direct time-domain sound data, divided to 130ms-long windows. According to the authors, CREPE's accuracy outperforms the state-of-the-art pYIN algorithm when tested on synthetic music data. However, no results are provided regarding CREPE performance on speech data.

A hybrid approach is taken by the authors of [10]. Sound is pre-conditioned with a PEFAC-like front end, operating on 40ms-long windows and then directed to a CNN. Finally, a post-processor is deployed, enforcing temporal continuity by the means of dynamic programming. This proposition is tested on speech data, where a performance gain is reported, when compared to PEFAC, an HMM-based method [11] and a deep neural network (DNN) based method [12]. The top score of 80% F_0estimation accuracy in +5 dB signal-to-noise ratio (SNR) condition is reported, allowing a ±5% tolerance.

978-1-6654-4112-4/21/$31.00 ©2021 IEEE

Another contribution utilizing a hybrid approach is HarFeature [13]. A lot of effort is put in the front-end for F_0 candidates selection. For both long-short-term spectrum and long-term subharmonic summation spectrum, a set of harmonic features is computed, e.g. harmonic energy ratio, subharmonic amplitude ratio, harmonic frequency deviation, odd to even harmonic energy ratio and ratio of identified harmonic partials. These extracted features constitute an input to a DNN classifier, followed by an HMM post-processor. Results are provided for two speech datasets: Keele pitch [14] and CSTR [15] and yield performance comparable to YIN algorithm for +20 dB SNR, with the peak of 95% F_0 estimation accuracy, assuming a ±5% tolerance.

Interestingly, more traditional machine learning solutions are proved not to be inferior to deep learning by the authors of [16] – which can be especially reassuring in case of the training data shortage. First, a set of 16 acoustic features is extracted from data (e.g. ACF or summation of the speech harmonics), which is then fed to a multilayer perceptron (MLP) or K-means classifier performing the voicing detection task. Finally, a simple final median filter is employed for F_0 estimation. Compared to RAPT and CREPE, this solution proved superior in voiced data classification.

Contrary to all previously presented pitch estimators, [17] stands out as the only one embodying the unsupervised learning paradigm. This is a significant distinction, as the availability of high-quality speech datasets with accurate and verified reference pitch annotations is a real problem. In terms of the architecture, this solution consists of 2 stages: the pre-processing stage, in which constant-Q transform of input is computed; and the CNN stage, outputting the final fundamental frequency prediction. Results are provided for singing data and show comparable quality to CREPE.

An odd one out is the method described in [18], consisting of a Bayesian pitch tracker. While arguably not a machine learning approach, its novelty makes it stand out from the conventional DSP category. Interestingly, on top of the Keele Pitch dataset, results are given also for Parkinson's disease speech dataset [19] and show performance gain in relation to most state-of-the-art pitch estimators, e.g. YIN, PEFAC and CREPE.

III. Data Creation and Feature Extraction

A. Dataset

As stressed before, a significant problem in supervised data-driven solutions development is the very core of it – accessibility of data. While a lot of speech datasets are available on various licensing terms (e.g. TIMIT [20], crowdsourced UK and Ireland English Dialect dataset [21] or GRID [22]), in case of pitch estimation a need arises to complement a speech dataset with an accurate set of annotations (e.g. PTDB-TUG [23] or Keele pitch). In order to work around this problem, the authors of this contribution chose to generate an artificial dataset, consisting of any number of synthetic vowel-like sounds associated with accurate pitch annotations with instantaneous temporal resolution.

The vowel synthesizer consists of a poly-harmonic signal generator in which harmonics' amplitudes are given by the following formula (1):

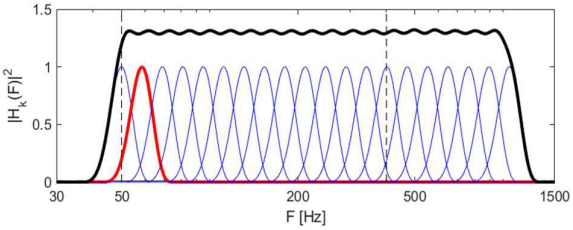


Fig. 1. CQT filterbanks.

$$A[k] = \{\sin[(2\pi \frac{k \cdot F_0}{F_s})]\}^{s^4} \quad (1)$$

where k is an index of the harmonic $F_k = k \cdot F_0$, ranging between a random F_0 to $\frac{1}{2}F_s$, s is the sampling rate and the resonance is established by the factor v randomly drawn from its range $v \in [0.15, 0.4]$, similarly to $s \in [1.4, 2.5]$. Finally, a synthetic vowel is synthesized by:

$$y[n] = \sum_{k=1}^{K} A[k] e^{j(k\varphi[n] + \varphi_0[k])}, \quad (2)$$

where

$$\varphi[n] = \frac{2\pi}{F_s} \sum_{m=0}^{n} F[m], \quad (3)$$

$F[\cdot]$ is a nonlinearly, randomly swept fundamental frequency and $\varphi_0[\cdot]$ is random. The model's fidelity is increased by an addition of random jitter and shimmer [24]. Finally, white gaussian noise is mixed to the signal at a controllable level, allowing for a precise SNR control.

Datasets created with this method were built of 500ms-long segments of synthetic vowel, interrupted with intervals of 0-200ms-long silence. F_0 ranged between 50Hz and 400Hz and two kinds of distributions were examined – logarithmic and linear.

B. Audio Front End

Data generated in the previous step was transformed using the Audio Front End (AFE) – a processing block whose goal is to extract features from raw audio data and prepare it for further processing. While in many embodiments feature extraction serves a purpose of reducing the amount of data to be processed, in this case an opposite is true. In this contribution, AFE is used to extract coarse estimates of instantaneous angular frequency of the signal's harmonics. For each sample in the signal, an instantaneous frequency and power is computed for each of 20 frequency sub-bands between 50Hz and 1050Hz – outputs of 20 Hilbert transform, constant-Q bandpass finite impulse response (FIR) filters designed using the kernel CQT method [25] with Blackman window (Fig. 1).

Within each sub-band, a complex mutual power operator (CMPO) is computed:

$$y_{\mathrm{CMPO}}[n] = x[n] \cdot x^*[n-1], \quad (4)$$

where $x[n]$ is the complex filter output and $x^*[n]$ its conjugate. Using CMPO, instantaneous power and frequency are obtained:

$$P_i[n] = |y_{\mathrm{CMPO}}[n]|, \quad (5)$$

$$F_i[n] = \arg(y_{\mathrm{CMPO}}[n]) \quad (6)$$

The output of AFE is not as accurate to be regarded a standalone pitch estimator, however it provides insight into the fundamental frequency and ratio between harmonics of vowels for the neural network to build its learning upon.

IV. Neural Network

The originally defined IFE method [26] leverages neural network as a core of its pitch estimation technique. In case of the presented work, the same structure has been utilized for further research and experiments. The chosen architecture consists of multilayer perceptron with two hidden layers with sizes of 500 and 1000 neurons [27]. It was determined via grid search cross-examinations, including experiments on number of hidden layers, neurons count, type of activation function, batch size and other NN hyperparameters. The input layer has been adjusted to fit data of 40 individual power and frequency features for each input sample provided by AFE.

A. Classification

The initially proposed classification mechanism categorized provided input data into 351 classes matching 1 Hz frequency bins in the range of 50-400 Hz. In the course of further experiments the authors decided to substitute linear classes with logarithmic distribution. As a result, the neural network has been re-trained with input data evenly distributed across newly defined 351 logarithmic classes.

Encouraged by promising results of classification methods adjustments, the authors focused on output bin sizes which initially divided 50-400 Hz range into 351 classes. Taking into consideration that majority of the proposed methods assume pre-defined classification accuracy margin [10, 13], the number of output classes in the depicted architecture has been reduced to 100. Fig. 2 depicts comparison of 351 and 100 bins classification for both linear and logarithmic distributions. Histogram presenting F_0 estimation for 351 linearly distributed classes shows significant variance for higher frequencies with some of the classes being very rarely selected. However, switching to 351 logarithmic classes lowers the variance, the most vivid improvement can be observed for both distributions with number of classes limited to 100.

B. Training & validation

Synthetic dataset has been generated for neural network training purposes. It consists of 60 million individual input samples normalized in the range of 0 to 1, which corresponds to approximately 2h 5min of continuous audio data, generated at 8kHz sampling rate. Equivalent synthetic data, with various SNR levels, have been used for neural network validation along with real speech data acquired from Keele pitch dataset. In addition to the described datasets, an enlarged training package of 240 million samples has been prepared to rule out scenario of data shortage. Both training datasets provided similar results which indicates that initially selected data size

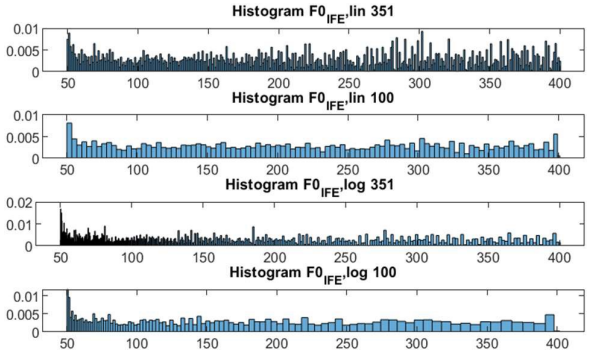


Fig. 2. Neural network classification per class.

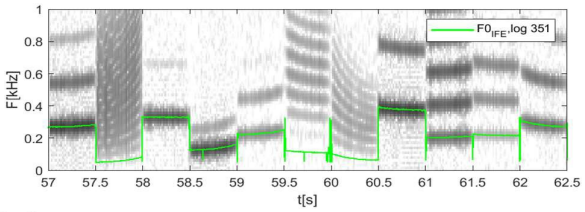


Fig. 3. Fragment of synthetic validation signal spectrogram with estimated F_0 for 351 classes distribured logarithmicaly.

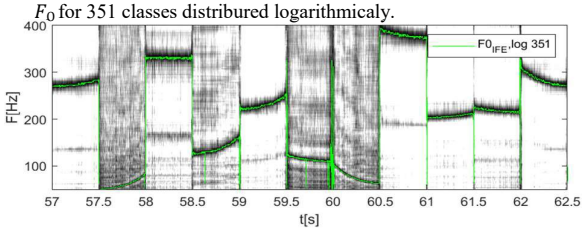


Fig. 4. Fragment of map of confidences per class for validation signal with estimated F_0 for 351 classes distribured logarithmicaly.

of 60 million samples was sufficient for neural network training.

Fig. 3 shows spectrogram of the fragment of a synthetic validation signal along with F_0 estimate. We can see here that for some segments F_0 aligns with the strongest signal component but there are also segments in which these components disappear in noise and the F_0 has to be determined from the distance between harmonic components. On the other hand, the neural network generates the map of confidences (Fig. 4) with the strongest component corresponding to F_0 . Moreover, the F_0 harmonics and subharmonics are also noticeable in this map, which in some cases, especially for lower SNRs, leads to estimation errors.

C. Parameterization

The presented network architecture has been trained through 10 epochs on each of the mentioned training datasets with batch size set to 64 samples. The model utilizes hyperbolic tangent which gave the best accuracy results in comparison to other popular activation functions. Categorical cross entropy has been chosen as a loss function along with Adam optimization algorithm. Classification leverages softmax function.

Implementation of the proposed architecture was written in Python programming language with Keras machine learning framework [28]. The same environment has been used for training, inference and optimalization purposes.

V. Results

Results discussed in subsections A and B were obtained for a synthetic validation dataset, generated with roughly 25dB of SNR.

A. Logarithmic and linear F_0 distribution

Fig. 5 presents how the F_0 estimation error is distributed depending on the actual instantaneous F_0 of the signal. As we can see, most of the estimates are in $\pm 1\%$ range (marked with red dashed lines) with many of the outliers kept in $\pm 5\%$ range (limited by the blue dashed lines). When comparing previously proposed network with linearly distributed 351 classes with reference methods (Fig. 5), the main difference can be seen in increased number of large outliers as well as slightly larger deviation of estimation errors for F_0 in range 50-80Hz. Change from linear to logarithmic distribution of classes limits these problems at the cost of slightly increased

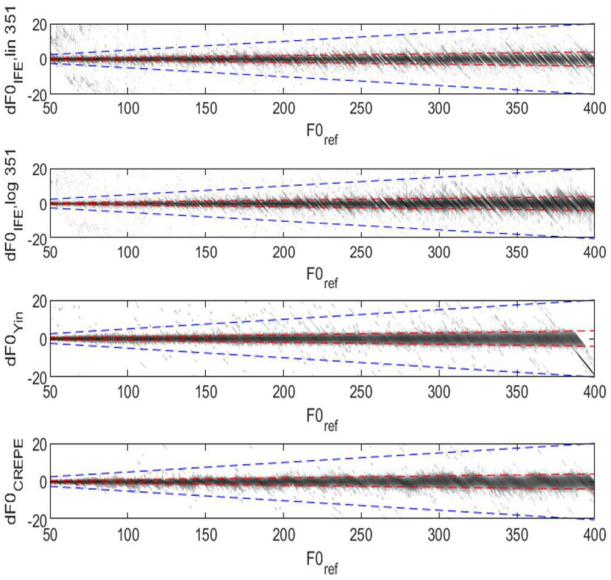


Fig. 5. Map of logarithm of counts of estimation error values observed for given reference F_0 of synthetic validation signal for neural networks with 351 classes and reference methods.

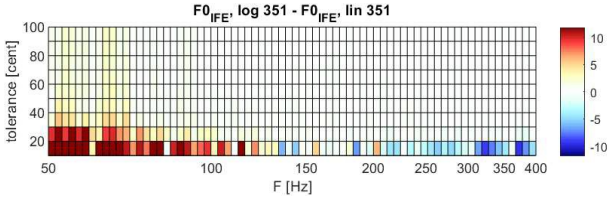


Fig. 6. Estimation accuracy improvement per frequency segment in % resulting from change from linear to logarithmic classes distribution.

deviation of estimation errors for high F_0 which can be seen in more detail in Fig. 6. The unit *cent* used in y-axis of the figure is a logarithmic pitch unit borrowed from the musical interval context, representing $1/100$ of a *semitone* – the smallest musical interval with a frequency ratio of $\sqrt[12]{2}$.

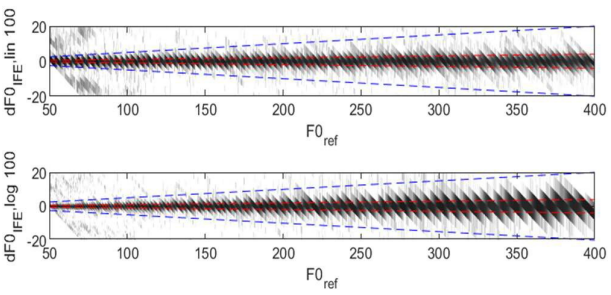


Fig. 7. Map of logarithm of counts of estimation error values observed for given reference F_0 of synthetic validation signal for 100 classes.

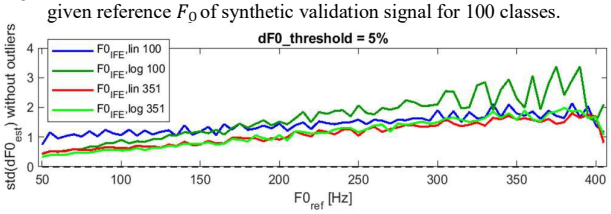


Fig. 8. Comparison of standard deviations of estimation errors evaluated for 5 Hz reference F_0 segments.

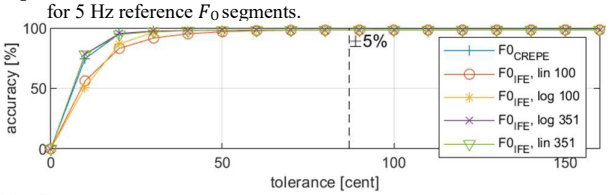
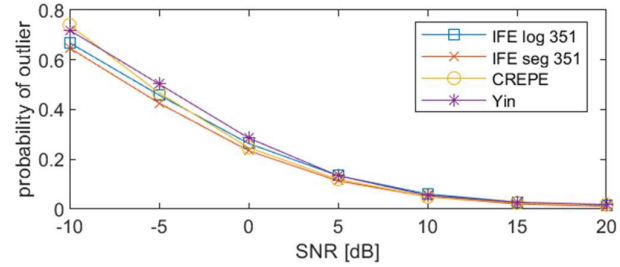


Fig. 9. F_0 estimation accuracy comparison for synthetic validation signal.

B. Performance in relation to the number of output classes

As mentioned before, the network with 351 classes seems not to be able to utilize all the classes equally for higher frequencies. This problem is slightly alleviated with logarithmic distribution of classes but larger improvement can be achieved with limiting number of classes to 100 (Fig. 7). This, however, results in deterioration of estimation error in consequence of coarser estimated F_0 quantization. Consequently, standard deviation of estimated F_0 increases, which limits accuracy depending on classes number and distribution (Fig. 8). Nonetheless, if larger tolerances (above 50 cents) can be accepted, such smaller network performance is similar to that of the larger network (Fig.9).



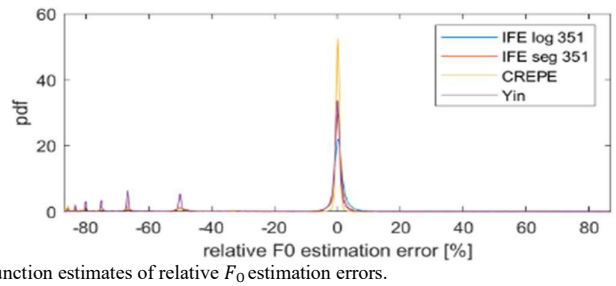
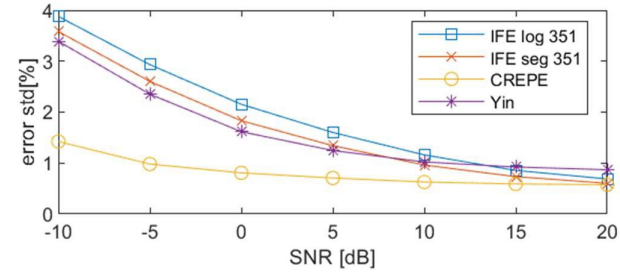


Fig. 10. Comparison of probability of outliers vs SNR and pobability density function estimates of relative F_0 estimation errors.



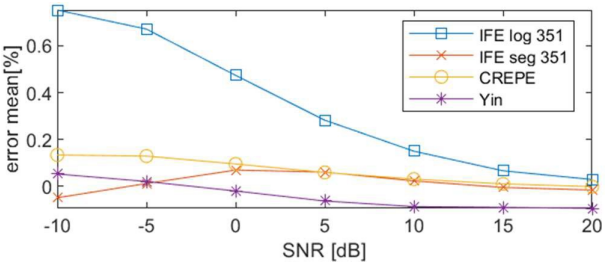


Fig. 11. Standard deviations and bias of relative estimation error.

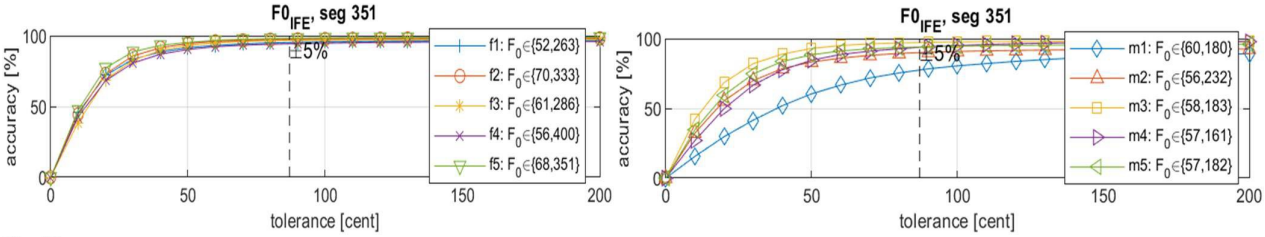


Fig. 12. F_0 estimation accuracy for female and male speech.

C. *Performance in relation to SNR*

Performance of the proposed estimator in presence of high gaussian noise has been measured for synthetic test signals and compared with CREPE and YIN methods. Here we compare the proposed network with 351 logarithmically distributed classes that provides F_0 estimated per each audio sample (IFE log 351 in figures). Since CREPE and YIN were configured to estimate F_0 in 10ms segments, the estimates postprocessed in 10ms segments have been additionally introduced (IFE seg 351 in figures). The F_0 estimated per segment is selected as an average of estimates what are in range of ±10% of median and is corrected with the help of std of averaged values to mitigate estimation bias.

In Fig. 10 we can see how the number of outliers increases for small SNRs. At 20dB all compared approaches demonstrate about 1.6% of outliers mostly located at original signal segments boundaries and with decreasing SNR the number of outliers increases up to 67% at SNR=−10dB with slight edge of our proposition over CREPE and YIN. From pdf of relative estimation error (Fig. 10) we can also see that with increased noise level the outliers tend to group around frequencies that are about 2, 3, 4 and times smaller than the actual F_0(-50%, -66,7%, -75% and -80%).

If large outliers (estimates with relative errors exceeding 10%) are discarded then the std of the relative error (Fig. 12) of the proposed method increases from 1% at 20dB to almost 4% at -10dB with some performance gain from post-processing in segment. This is similar to YIN that performs better in this respect for lower SNR but achieves worse results for high SNRs. The best results are obtained with CREPE and the proposed method performing similarly for high SNR. If relative estimation error bias (mean) is compared (Fig. 11), then the raw proposed method demonstrates larger bias which could be effectively decreased to the level of other methods with post-processing in 10ms segments.

D. *Performance for speech signals*

Since the goal of the proposed network training is to estimate F_0 for speech signals, another test has been done on Keele pitch speech dataset. As can be seen in Fig. 12, better accuracy is achieved for female speech while male speech with lower F_0 pose more of a challenge, particularly the m1 speaker.

When we compare our previous network with current proposition with logarithmic classes distribution and per segment post-processing (Fig. 13) then we can notice small improvement in accuracy for low tolerances (about 50 cents) for female speech and significant improvement of accuracy for all male speech samples. Moreover, for tested dataset the proposed method performed better than CREPE (Fig. 14), especially for male speakers.

VI. Post Processing with Hidden Markov Model

It is apparent from the results analysis that one of the primary types of error in pitch estimation is a phenomenon of favoring a harmonic over the fundamental. This may be due to a vowel resonance resulting in higher harmonic amplitude, masking by noise or other effects. In case of female speech this problem is marginal. For example, in the map of network confidences obtained for female speech with F_0 around 200Hz (Fig 15) we can notice that for voiced segments there is very little unambiguity between F_0 and its harmonic. On the other hand, in case of m1 male speech map of network confidences (Fig. 16) shows similar confidence levels for F_0 and its harmonic, which results in decision switching between those two values. Tests performed for speech and synthetic signals show that for the proposed method as well as for

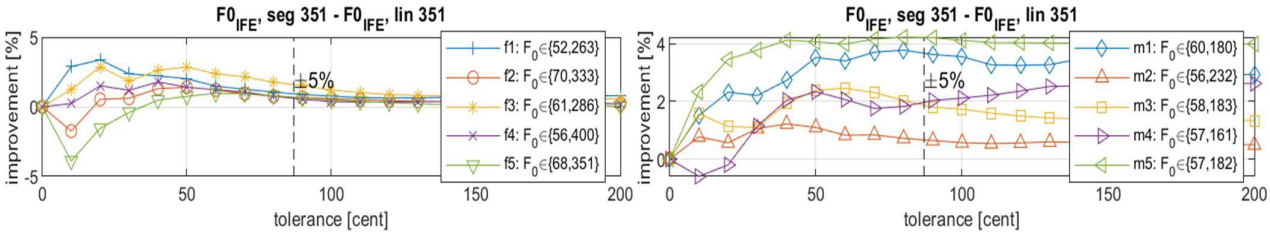


Fig. 13. F_0 estimation accuracy improvement for female and male speech resulting from change to logarithmically distributed classes.

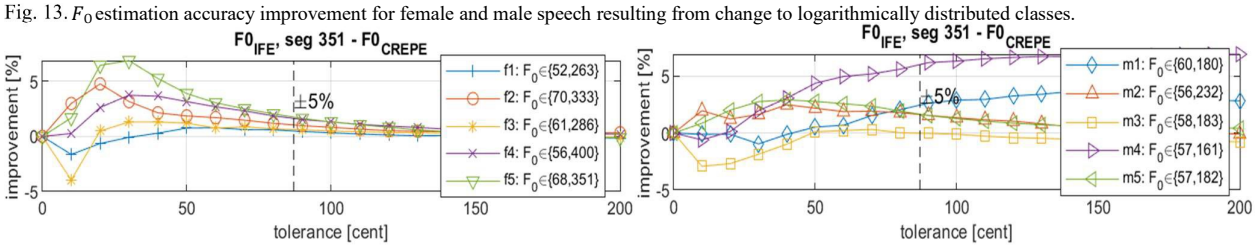


Fig. 11. F_0 estimation accuracy improvement for female and male speech for the proposed method in comparison to CREPE.

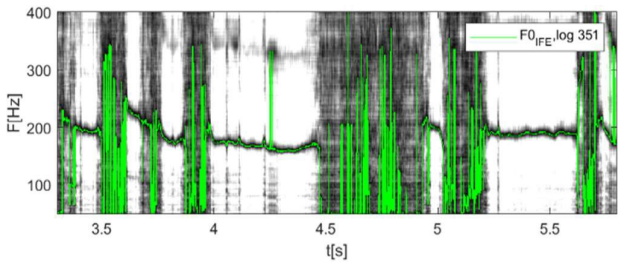


Fig. 15. Map of confidences per class of femal speech signal f5.

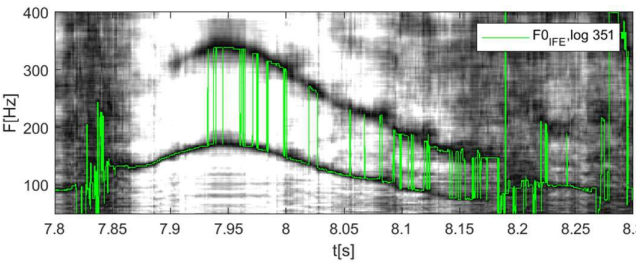


Fig. 16. Fragment of map of confidences per class for male speech signal m1.

competitive methods, mistaking F_0 with its harmonics is the most common problem.

In cases where harmonic candidate is picked over the fundamental throughout only a portion of the vowel's duration, the confidence map (Fig. 16) shows also high confidence for actual F_0. This type of error can be handled with some kind of temporal continuity enforcement. As noted during the literature review, a common method used in solving this problem is incorporation of a hidden Markov model (HMM).

In principle, HMM is a probabilistic model of a given time series and can be successfully used in plethora of applications. For pitch estimation, its purpose is to serve as a pitch tracker – a model responsible for determining what is the likelihood of certain pitch being followed by a given one. To achieve that goal, HMM must be armed with an emission matrix – a square matrix of probabilities of subsequence of each pair of allowed pitches. In this case, the emission matrix was extracted from MATLAB built-in set of examples [29], which in turn was derived from PTDB-TUG dataset [23].

In this effort, only linearly-spaced NN output of 351 pitch classes confidences was considered. During the algorithm run, first a set of 10 highest confidence candidates was picked from all 351 candidates. Then, for each of those candidates, its confidence was weighted by a matching factor in the emission matrix. Afterwards, a backward run through the confidence matrix was performed to pick the highest confidence pitch contour. In Fig. 17, a sample excerpt spectrogram of Keele pitch dataset is shown with IFE predictions without (white) and with (red) HMM post-processing. It shows a promising degree of improvement, however further refinement and enabling logarithmic output processing is necessary in the future.

VII. Conclusions and Future Directions

The presented work provides a deep analysis and refinement of novel pitch estimation method called IFE. The main areas of interest covered in this paper are as follows:

- Influence of linear and logarithmic class distribution and number of classes on the neural network ability to estimate pitch.
- Impact of various SNR levels on IFE method accuracy.
- Analysis of outlier errors and suggestion of potential post-processing to overcome them.
- Comparison of IFE with current state-of-the-art traditional and machine learning based methods as YIN and CREPE.

The authors scrupulously examined and validated IFE method with various variants of input data and classification aspects. The conduced research and improvements allowed for achieving convergent results with other state-of-the art methods as CREPE and YIN. It is crucial to remark that contrary to previously mentioned techniques, IFE is instantaneous – it does not require any time segmentation of the input data, resulting in a sample-level temporal resolution of pitch estimation. However, if this degree of precision is not required, an additional post-processing could be implemented to refine the IFE output with some time-windowing approach, e.g. a median filter.

An additional distinguishing factor of the IFE approach is generation of training dataset. The presented results of F_0 estimation for speech signals were achieved based on NN training with solely synthetic data. Such approach eliminates issues with establishing ground truth F_0 values for input signals during training step and allows for efficient preparation of large datasets required by neural network models.

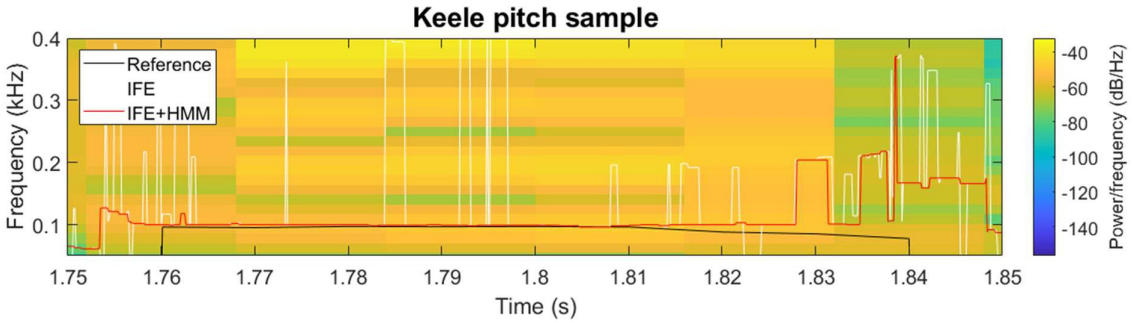


Fig. 17. An example of HMM post processing effect.

Despite satisfactory results, the authors still see plenty of room for IFE method improvements. More sophisticated post-processing methods could create significant opportunity for limitation of the number of outliers in the output estimation. Apart from this, there is still a need for robust grid search of optimal structure and parametrization of the model. Finally, special attention should be put on Audio Front End calibration to ensure that feature provided to the trained model are sufficient for optimal classification results.

REFERENCES

[1] R. L. Miller and E. S. Weibel. "Measurement of the fundamental period of speech using a delay line," in The Journal of the Acoustical Society of America, vol. 28, no. 4, pp. 761-761, 1956.

[2] L. Rabiner, M. Cheng, A. Rosenberg and C. McGonegal, "A comparative performance study of several pitch detection algorithms," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, no. 5, pp. 399-418, October 1976, doi: 10.1109/TASSP.1976.1162846.

[3] B. Bechtold, "Pitch of Voiced Speech in the Short-Time Fourier Transform: Algorithms, Ground Truths, and Evaluation Methods," PhD Dissertation at Jade Hochschule & Carl von Ossietzky Universität Oldenburg, Germany, 2020.

[4] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," Speech coding and synthesis, W. B. Kleijn and K. K. Palatal, Elsevier Science Inc., USA, pp. 495-518, 1995.

[5] A. De Cheveigné and H. Kawahara. "YIN, a fundamental frequency estimator for speech and music," The Journal of the Acoustical Society of America, vol. 111, no. 4, pp. 1917-1930, 2002.

[6] C. O'Reilly, N. Marples, D. Kelly and N. Harte, "YIN-Bird: Improved Pitch Tracking for Bird Vocalisations", Interspeech 2016, 2016.

[7] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions", 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.

[8] S. Gonzalez and M. Brookes, "PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 2, pp. 518-530, 2014.

[9] J. W. Kim, J. Salamon, P. Li and J. P. Bello, "Crepe: A Convolutional Representation for Pitch Estimation," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 161-165, doi: 10.1109/ICASSP.2018.8461329.

[10] H. Su, H. Zhang, X. Zhang and G. Gao, "Convolutional neural network for robust pitch determination," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 579-583, doi: 10.1109/ICASSP.2016.7471741.

[11] Z. Jin and D. Wang, "HMM-Based Multipitch Tracking for Noisy and Reverberant Speech," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1091-1102, July 2011, doi: 10.1109/TASL.2010.2077280.

[12] K. Han and D. Wang, "Neural networks for supervised pitch tracking in noise," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014, pp. 1488-1492, doi: 10.1109/ICASSP.2014.6853845.

[13] D. Wang, C. Yu and J. H. L. Hansen, "Robust Harmonic Features for Classification-Based Pitch Estimation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 5, pp. 952-964, May 2017, doi: 10.1109/TASLP.2017.2667879.

[14] F. Plante, G.F Meyer and W.A Ainsworth, "A pitch extraction reference database", In Fourth European Conference on Speech Communication and Technology EUROSPEECH-1995, 837-840, 1995.

[15] P.C. Bagshaw, S.M. Hiller and M.A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching" in Proceedings of the 3rd European Conference on Speech Communication and Technology, 1993.

[16] T. Drugman, G. Huybrechts, V. Klimkov and A. Moinet, "Traditional Machine Learning for Pitch Detection," in IEEE Signal Processing Letters, vol. 25, no. 11, pp. 1745-1749, Nov. 2018, doi: 10.1109/LSP.2018.2874155.

[17] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi and M. Velimirović, "Pitch Estimation Via Self-Supervision," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 3527-3531, doi: 10.1109/ICASSP40776.2020.9053798.

[18] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little and M. G. Christensen, "Robust Bayesian Pitch Tracking Based on the Harmonic Model," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, pp. 1737-1751, Nov. 2019, doi: 10.1109/TASLP.2019.2930917.

[19] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive Kalman filtering," The Journal of the Acoustical Society of America, vol. 135, no. 5, pp. 2885–2901, 2014, https://doi.org/10.1121/1.4870484.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1". NASA STI/Recon technical report n, 93, 27403, 1993.

[21] Demirsahin, I., Kjartansson, O., Gutkin, A. and Rivera, C., 2020. Open-source Multi-speaker Corpora of the English Accents in the British Isles. Proceedings of The 12th Language Resources and Evaluation Conference (LREC), [online] pp.6532--6541. Available at: <https://www.aclweb.org/anthology/2020.lrec-1.804> [Accessed 30-Apr-2021].

[22] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition". The Journal of the Acoustical Society of America, vol. 120, no. 5, pp. 2421-2424, 2006.

[23] G. Pirker, M. Wohlmayr, S. Petrik and F. Pernkopf, "A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario", 12th Annual Conference of the International Speech Communication Association, pp. 1509-1512, 2011.

[24] J. P. Teixeira, C. Oliveira and C. Lopes, "Vocal acoustic analysis–jitter, shimmer and HNR parameters," Procedia Technology, 9, pp. 1112-1122, 2013.

[25] C. Schörkhuber, and A. Klapuri. "Constant-Q transform toolbox for music processing." 7th Sound and Music Computing Conference, Barcelona, Spain. 2010.

[26] M. Blok, J. Banaś and M. Pietrołaj, "Neural Network Classifier for Instantaneous Pitch Estimation", unpublished.

[27] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," Psychological review, vol. 65, no. 6, pp. 386, 1958.

[28] Team, K., 2021. Keras: the Python deep learning API. [online] Keras.io. Available at: <https://keras.io/> [Accessed 30-Apr-2021]..

[29] "Pitch Tracking Using Multiple Pitch Estimations and HMM- MATLAB & Simulink- MathWorks Nordic", Se.mathworks.com, 2021. [Online]. Available: https://se.mathworks.com/help/audio/ug/pitch-tracking-using-multiple-pitch-estimations-and-hmm.html. [Accessed: 30-Apr-2021].