TOWARDS CONFIDENCE MEASURES ON FUNDAMENTAL FREQUENCY ESTIMATIONS

Boyuan Deng[1−4]*, Denis Jouvet*[1−3]*, Yves Laprie*[1−3]*, Ingmar Steiner*[5,6]*, Aghilas Sini*[1−3]

Speech Group, LORIA
[1]Inria, Villers-lès-Nancy, F-54600, France
[2]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[3]CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[4]Tencent [5]Saarland University, Germany [6]DFKI GmbH

bryandeng@tencent.com;{denis.jouvet,yves.laprie,aghilas.sini}@loria.fr;ingmar.steiner@dfki.de

ABSTRACT

The fundamental frequency is one of the prosodic parameters, and many algorithms have been developed for estimating the fundamental frequency of speech signals. Most of them provide good results on good quality speech signals, but their performance degrades when dealing with noisy signals. Moreover, although some provide a probability for the voicing decision, none of them indicate how reliable the estimated fundamental frequency is. In this paper, we investigate the computation of a confidence (or reliability) measure on the estimated fundamental frequency values. A neural network based approach is proposed for computing the posterior probability that the estimated fundamental frequency is correct. Experiments are conducted on the PTDB-TUG pitch-tracking database, using three fundamental frequency estimation algorithms.

***Index Terms*—** Pitch, fundamental frequency, estimation, confidence measures.

1. INTRODUCTION

Fundamental frequency (also called F0) is one of the prosodic parameters, along with energy and phone duration. Prosody reflects various characteristics of the speaker and/or of the utterance, such as the emotional state of the speaker, the modality of the utterance (question, statement, ...), as well as emphasis on some words. Prosodic features are also involved in marking lexical stress. Hence, the determination of the fundamental frequency is required in many applications including computer assisted language learning, and in the development of automatic speech synthesis systems, especially for expressive speech synthesis.

Numerous algorithms have been developed in the past for computing the fundamental frequency of speech signals. On good quality speech signals, the various algorithms work quite well. However there exists some mis-detections (F0 value associated with an unvoiced frame, or F0 not detected for a voiced frame), as well as wrong estimations of the F0 values. Moreover, errors become more frequent on lower quality speech data and on noisy speech data; sometimes, intrinsic voicing quality is also a problem. All algorithms provide a voiced/unvoiced decision, as well as F0 values on voiced portions of the signal, but, although a few algorithms provides a probability for the voicing decision, none of them indicate if the estimated F0 values are reliable or not.

Some F0 estimation algorithms operate in the time domain relying on autocorrelation methods (e.g., [1]), while some other methods operate in the frequency domain, for example the spectral comb approach [2] and the Sawtooth Waveform Inspired Pitch Estimator (SWIPE) method [3]. Many variants have been developed, and some of them combine processing in the time and in the frequency domains (e.g., [4]). As this study does not aim to compare various F0 estimation algorithms, we do not introduce nor discuss other F0 detection algorithms. Indeed, the goal of this study is to investigate the computation of confidence measures on the estimated F0 values, and in this preliminary study, we limit our investigation to three F0 estimation algorithms (based on [1], [2] and [3]) as implemented in the JSnoori toolkit [5].

Confidence measures have been studied for a long time in automatic speech recognition. As detailed in [6], confidence measures can be based on a combination of predictor features that have different distribution between correctly and incorrectly recognized words, or based on hypothesis testing (i.e., utterance verification) or defined as a posterior probability of the word. The later is currently the most often used in speech recognition systems. In this study we choose to rely on the posterior probability for defining a confidence measure on the estimated pitch values. For computing such posterior probability we rely on a neural network based classifier, which, when trained with a squared error or a cross-entropy cost function provides Bayesian probabilities [7, 8].

The goal of the proposed study is to develop and evaluate an approach that estimates the reliability of the fundamental frequency values provided by a F0 detection algorithm. As

a support for the evaluations we rely on a reference speech database, namely the Graz University of Technology Pitch-Tracking Database (PTDB-TUG) [9]. This data is then artificially corrupted by adding different types of noise signals (from the NOISEX-92 corpus [10]) at various signal to noise ratios. The analysis of the F0 results provides insight into the robustness of the various algorithms. Various features (computed on the signal and/or used in the estimation of the fundamental frequency) are also extracted and are given as input to the neural network for computing the posterior probability that the F0 estimate is correct.

The paper is organized as follows. Section 2 provides an overview of the experimental framework. Section 3 gives an insight into the performance of the various F0 detection algorithms. Section 4 presents the main contribution of the paper, i.e., the computation of confidence measures on the estimated F0 values, and analyzes the performance of the corresponding classifier. A conclusion ends the paper.

2. EXPERIMENTAL FRAMEWORK

A speech database with reference F0 values is used, and in order to study the behavior of the F0 detection algorithms with respect to the speech signal quality, noise recordings are added to the clean speech data at various signal to noise ratios.

2.1. F0 detection algorithms

Keeping in mind that our goal is not to compare in details the performance of various F0 detection algorithms, but rather to elaborate a method for computing confidence measures indicating the reliability of the estimated F0 values, we consider the following F0 detection algorithms for this study.

The first algorithm is based on Martin's algorithm [2]. It is a spectral based method which detects fundamental frequency by maximizing the intercorrelation between the power spectrum and a spectral comb. The implementation used is an enhanced version which takes into account energy histograms and voicing jumps in post-processing steps, and which was optimized for processing clean speech signals.

The second is Sawtooth Waveform Inspired Pitch Estimator (SWIPE) [3], which takes into account the fact that peaks in the spectrum are separated by valleys; hence the algorithm finds the F0 candidates by maximizing the average peak-to-valley energy distances. The algorithm also includes blurring and weighting of the harmonics, as well as spectrum warping.

The third algorithm is the YIN approach [1] which is basically a correlation-like method in the time domain. YIN finds the fundamental period by minimizing a criterion based on the time difference function which is much less sensitive to amplitude changes than the auto-correlation function. Some post-processing is also carried out for improved performance.

Note that, in the following experiments, there was no tuning of the configuration parameters of each algorithm.

2.2. Speech corpus

The Graz University of Technology Pitch-Tracking Database (PTDB-TUG) is used for the reported experiments. This database contains microphone and laryngograph signals of 20 English native speakers reading out sentences from the TIMIT corpus [11]. The speakers are gender balanced (10 female, 10 male) and their ages vary from 22 to 48 years. Overall, a total of 4720 sentences were recorded.

Both microphone and laryngograph signals were recorded at 48 kHz sampling rate, with 16 bit signed PCM encoding. The database also provides reference pitch trajectories that were extracted from the laryngograph signals using the Robust Algorithm for Pitch Tracking (RAPT) [12]. Reference F0 values are provided at 10 ms intervals.

The speech signal was downsampled to 16 kHz prior to computing the F0 values every 10 ms with the three algorithms described in Section 2.1.

2.3. Noise data

As the PTDB-TUG corpus contains only clean speech signals recorded in a studio environment, in most of the following experiments, some noise is added to the clean speech signals. Noise recordings are taken from the NOISEX-92 corpus [10], which is often used in the field of automatic speech recognition. The following noise types have been considered: *babble* (people speaking in a canteen), *factory1* (sound recorded near plate-cutting and electrical welding equipment in a factory), *factory2* (sound recorded in a car production hall), *pink* (acquired by sampling a high-quality analog noise generator (Wandel & Goltermann), yielding equal energy per 1/3 octave), and *white* (acquired by sampling the same analog noise generator, with equal energy per Hz bandwidth).

All these noise signals were downsampled to 16 kHz before being added to the clean speech signal at different signal-to-noise ratios (SNRs) using the Filtering and Noise adding Tool (FaNT) [13, 14] toolkit. This tool has been used for the creation of artificially distorted speech data inside the evaluations of the ETSI working group Aurora [15]. It allows adding noise at a given signal-to-noise ratio, and it was used here for adding the noise recordings at 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB SNR levels.

3. F0 ESTIMATION PERFORMANCE

For each algorithm, the performance is evaluated by comparing the estimated F0 values to the reference values provided with the database. According to the evaluation terminology mentioned in [16] and [17], we consider mainly the *F0 frame error* which is the proportion of frames for which either a *voicing decision error* or a *gross pitch error* is observed. A voicing decision error is observed when a voiced frame is detected as unvoiced, or when an unvoiced frame is detected as voiced. A gross pitch error corresponds to a voiced frame

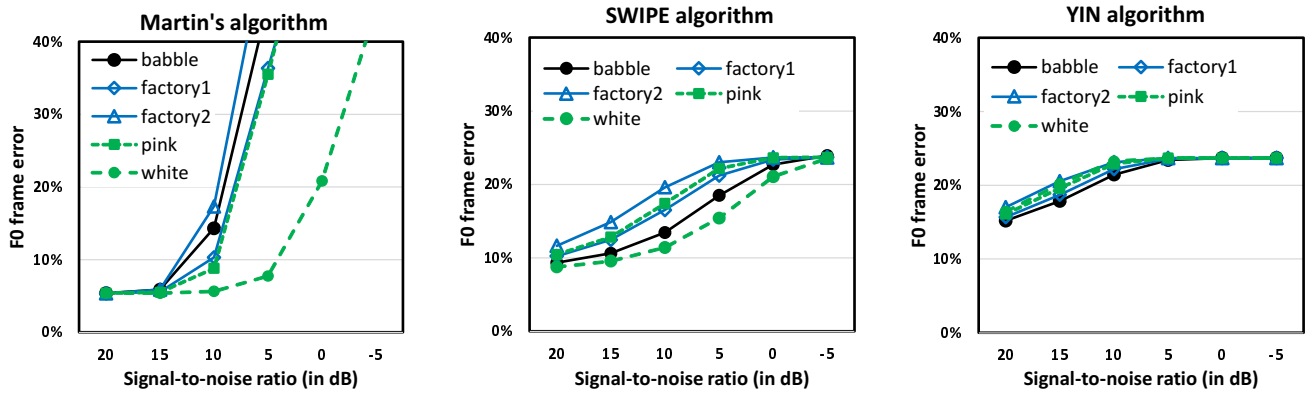

Fig. 1. F0 estimation errors on the reference data corrupted by various types of noise at several signal-to-noise ratios.

which is detected as voiced, but for which the estimated F0 value differs by more than 20 % from the reference F0 value (this includes errors for which the estimated F0 is doubled or halved).

In this preliminary study, the performance is evaluated on one fifth of the data (a random subset containing data from each of the 20 speakers). This subset is artificially corrupted by adding each of the five types of noise (babble, factory1, factory2, pink, and white) at various SNR levels. Figure 1 reports the F0 frame error with respect to the SNR level for the three F0 detection algorithms and the five types of noise.

As can be expected, the F0 frame error increases when the quality of the signal is degraded (i.e., when the SNR gets lower), and for every type of noise. However the degradation differs among the three algorithms. For the 20 dB and 15 dB SNR, the Martin based approach provides the best results, whereas its F0 frame error continues to rise for lower SNR. For the SWIPE and YIN approaches, the F0 frame error levels off at about 23 %, which corresponds to the percentage of voiced frames in the data; i.e., for low SNR levels, these algorithms tends to consider each frame as unvoiced.

4. CONFIDENCE MEASURES ON F0 ESTIMATES

This section details the proposed approach for computing a confidence measure on the estimated F0 values. Let's recall that although some F0 detection algorithms produce a probability of voicing for each frame, to the best of our knowledge, no F0 detection algorithm provides a confidence measure associated with the estimated F0 values.

We treat the prediction of correctness of the estimated F0 values as a classification problem (estimated F0 value correct vs. incorrect), using neural network based classifiers. According to [7] and [8], when trained with a cross-entropy objective function, the neural network output is a Bayesian probability. Hence the output of the neural network provides the posterior probabilty that the F0 estimated value is correct.

4.1. Neural network based classifiers

Two types of neural networks are considered in the experiments: a multilayer perceptron (MLP) and a recurrent network with long short-term memory (LSTM) cells [18]. Rectified linear units (ReLU) [19] are used for nonlinearities in the hidden layers; and a sigmoid activation is used for the output cell. During training, the cross entropy cost function is used, as well as dropout [20] to prevent overfitting.

For each F0 detection algorithm, the 59 frame features used as input to the neural network for computing the confidence measure are the energy of the frame, three F0 candidates and associated scores resulting from internal computations of the considered F0 detection algorithm, 40 cepstral coefficients, and 12 Mel-frequency cepstral coefficients.

The MLP network was designed with two fully connected hidden layers (each with 128 units) with rectified linear units. The LSTM network was designed with two LSTM recurrent layers (each with 128 units). In both cases, the output layer has a unique output unit with a sigmoid activation.

4.2. Classifier training and global performance

As the F0 detection algorithms works quite well on clean data, the amount of incorrect estimated values (i.e., negative examples) on clean data is too small for a proper training of the neural network. Hence examples associated with the lowest SNR (-5 dB), for which the F0 frame error rate is very high (thus leading to many negative examples), were also included in the training set. Using both the clean and the -5 dB SNR data helps balance the amount of positive and negative examples. Overall this leads to about 7 million samples for each F0 estimation algorithm.

In the first set of experiments, this data set was randomly split into two parts: 80 % of the data for training and 20 % for evaluation tests. Note that for the LSTM models, in order to maintain the temporal order of the samples, and avoid invalid sequences across the boundaries of audio files, the split

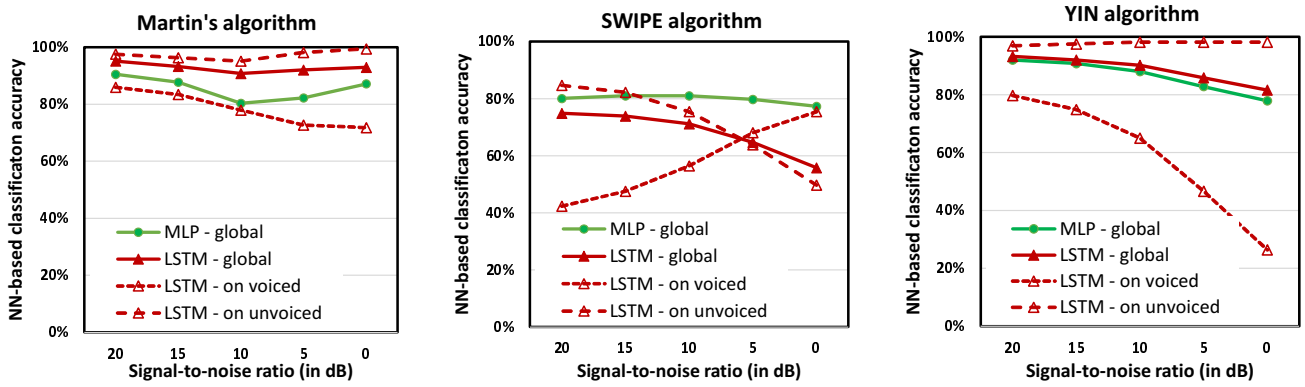

Fig. 2. Classification accuracy of the F0 values estimated on the reference data corrupted by various types of noise at several signal-to-noise ratios. Global classification rate at top, and classification rates on voiced and unvoiced frames at bottom.

is done at the file level. During the training process, 20 % of the training data is used as a development set. The models are trained using stochastic gradient descent on mini batches with cross entropy as objective cost function.

Table 1. Classification accuracy (in percent) of the MLP and LSTM neural networks for the F0 values estimated by the three F0 detection algorithms.

Model	Martin's	SWIPE	YIN
MLP	91.6	87.4	86.3
LSTM (seq. length 2)	92.6	88.0	89.3
LSTM (seq. length 3)	93.2	88.3	90.1
LSTM (seq. length 4)	93.4	88.6	90.2

Table 1 reports the classification accuracy achieved on the test set for the three F0 detection algorithms and the various modeling approaches. For the results reported in this table, the classification threshold was arbitrarily set to 0.5 (i.e., if the neural network output is greater than 0.5, the corresponding estimated F0 value is considered to be correct, and incorrect otherwise). So far, no study has been conducted to analyze the impact of this classification threshold, nor for optimizing it. The results show that the LSTM models provide much better classification accuracy than the MLP model, and this result holds for all three F0 estimation algorithms. With respect to the LSTM approaches, increasing the input sequence length (more time steps) leads to better classification accuracy.

4.3. Discussion

To further analyze the performance of the proposed models, Figure 2 shows the classification accuracy with respect to the various SNR levels. In two cases (Martin's and YIN algorithms), the LSTM approach leads to better classification performance than the MLP approach (solid lines, red triangle vs. green circle). Although training was carried out using only clean data and -5 dB noisy data, good classification results are achieved on this test set ranging from 20 dB to 0 dB SNR (that is, for SNR levels not present in the training set). We also display the classification accuracy for voiced and unvoiced frames; for the sake of legibility, only the LSTM results are detailed with respect to voiced (red short dash curves) and unvoiced (red long dash curves) frames. In most cases, the classification accuracy is much higher on unvoiced frames than on voiced frames (except for SWIPE at very low SNR). This may be due to the fact that there are many more unvoiced than voiced frames in the data (about 77 % unvoiced vs. 23 % voiced frames), which leads to rather high classification rates even for medium quality speech (15 dB and 10 dB SNR). We expect that increasing the training set would further improve the performance of the approach.

5. CONCLUSION

This paper has presented a neural network based approach for computing a confidence measure on F0 values estimated by F0 detection algorithms. With a classification threshold of 0.5, the proposed approach leads to high classification accuracy on the estimated F0 values, for various SNR levels.

This preliminary study on the computation of confidence measures on the F0 estimated values, along with the fact that various algorithms make different types of errors, opens the way to optimally combining several F0 detection algorithm to achieve a precise F0 detection that would be robust to speech signal quality.

6. ACKNOWLEDGMENT

This work was carried out in the framework of the Prosod-Corpus operation supported by the CPER LCHN (*Contrat Plan Etat Région "Langues, Connaissances et Humanités Numériques"*).

7. REFERENCES

[1] Alain de Cheveigné and Hideki Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[2] Philippe Martin, "Comparison of pitch detection by cepstrum and spectral comb analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1982, pp. 180–183.

[3] Arturo Camacho, *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*, Ph.D. thesis, University of Florida, 2007.

[4] Alexander Sorin, Tenkasi Ramabadran, Dan Chazan, Ron Hoory, Michael McLaughlin, David Pearce, Fan CR Wang, and Yaxin Zhang, "The ETSI extended distributed speech recognition (DSR) standards: client side processing and tonal language recognition evaluation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. I, pp. 129–132.

[5] "JSnoori," https://raweb.inria.fr/rapportsactivite/RA2015/multispeech/uid43.html.

[6] Hui Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, no. 4, pp. 455–470, 2005.

[7] Michael D. Richard and Richard P. Lippmann, "Neural network classifiers estimate Bayesian a posteriori probabilities," *Neural Computation*, vol. 3, no. 4, pp. 461–483, 1991.

[8] Guoqiang Peter Zhang, "Neural networks for classification: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 4, pp. 451–462, 2000.

[9] Gregor Pirker, Michael Wohlmayr, Stefan Petrik, and Franz Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Interspeech*, 2011, pp. 1509–1512.

[10] Andrew Varga and Herman J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[11] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathon G. Fiscus, and David S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[12] David Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. Bastiaan Kleijn and Kuldip K. Paliwal, Eds., pp. 495–518. Elsevier, 1995.

[13] "FaNT – filtering and noise adding tool," http://dnt.kr.hs-niederrhein.de/index964b.html.

[14] Hans-Günter Hirsch, "FaNT – filtering and noise adding tool," Tech. Rep., Hochschule Niederrhein, 2005.

[15] Hans-Günter Hirsch and David Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000 – Automatic Speech Recognition: Challenges for the new Millenium*, 2000.

[16] Thomas Drugman and Abeer Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Interspeech*, 2011, pp. 1973–1976.

[17] Onur Babacan, Thomas Drugman, Nicolas d'Alessandro, Nathalie Henrich, and Thierry Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 7815–7819.

[18] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] Vinod Nair and Geoffrey E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *International Conference on Machine Learning*, 2010, pp. 807–814.

[20] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.