A comparison of state-of-the-art speech fundamental frequency estimators in noisy and reverberant environments

Robert Rehr, Martin Krawczyk, Timo Gerkmann

Speech Signal Processing Group, Dept. Med. Physics & Acoustics, Cluster of Excellence “Hearing4all”, Universität Oldenburg

firstname.lastname@uni-oldenburg.de

Introduction

Many applications in speech signal processing rely on an accurate estimate of the fundamental frequency. This information may be used e. g. for coding speech signals efficiently or for enhancing noisy speech [1]. Thus, algorithms are required that are able to estimate the fundamental frequency reliably also in noisy and reverberant environments. We compare a wide range of methods for estimating the fundamental frequency which employ different features in their estimation procedure. These algorithms are compared with respect to their estimation accuracy in noisy and reverberant environments. Additionally, the computational complexity of these algorithms is considered. Experiments are conducted on speech utterances which are artificially corrupted by real-world background noises and reverberated using measured room impulse responses. We evaluate the algorithms in terms of the gross error rate (GER), which measures large deviations of the estimate from the ground truth, such as doubling and halving errors. The estimators' accuracy is further evaluated using the root-mean-square error (RMSE).

Fundamental frequency estimation

Roughly, speech sounds can be grouped into two types which are known as *voiced* and *unvoiced*. Unvoiced sounds reveal a rather noisy character caused by the air flow passing constrictions in the vocal tract, whereas voiced sounds are created by the vibration of the vocal folds. The latter results in a periodic and harmonic signal which is mainly characterized by the fundamental frequency. As the fundamental frequency is an important parameter for describing speech signals, it is utilized in many speech signal processing applications.

The estimation of the fundamental frequency has been investigated for several decades, e. g. [2]. Still, new methods are proposed, e. g. [3], in order to improve the robustness in noisy and reverberant environments. In our comparison, we include the methods described in [3–6] and a cepstral peak picking algorithm which is based on the work in [2]. A brief description of these estimators is given below.

YIN [4] exploits the periodicity in the time-domain and employs a modified version of the autocorrelation function for obtaining the fundamental frequency in short time-frames.

The pitch estimation filter with amplitude compression (PEFAC) [3] operates in the short-time frequency domain where a logarithmic frequency warping is employed. In the logarithmic frequency domain, a harmonic spectrum is represented as a fixed pattern of peaks which is shifted depending on the underlying fundamental frequency.

The cepstral peak picking algorithm proposed in [2] computes the cepstra of short-time frames. In the cepstral domain, the period length of the underlying fundamental frequency is represented by a peak which needs to be identified.

The non-linear least squares (NLS) algorithm [5, 6] uses the harmonic model to derive a likelihood function for the fundamental frequency. The maximum of the likelihood function is used to determine the fundamental frequency.

Evaluation method

For our comparison, we use sentences from the TIMIT core set [7] for which an annotation of the fundamental frequency has been provided by [8]. The utterances have been corrupted by white and babble noise taken from the NOISEX-92 [9] database at signal-to-noise ratios (SNRs) ranging from -10 to 20 dB in 5 dB steps. Additionally, we reverberate the speech signals using two room impulse responses taken from the Aachen Impulse response corpus [10]. The reverberation times T of the two selected rooms are $T_1 \approx 300$ ms and $T_2 \approx 800$ ms with direct-to-reverberant ratios (DRRs) of $\mathrm{DRR}_1 = 0.3$ dB and $\mathrm{DRR}_2 = -2.4$ dB, respectively.

The performance of the fundamental frequency estimators is evaluated using the GER, which accounts for large pitch deviations, and the RMSE, which quantifies fine pitch deviations. The GER is the mean over all voiced segments of the gross errors $G(\hat{f}_0[\ell])$, where $f_0[\ell]$ and $\hat{f}_0[\ell]$ denote the estimated and the annotated fundamental frequency of the ℓth frame, respectively. The gross error $G(\cdot)$ is given as

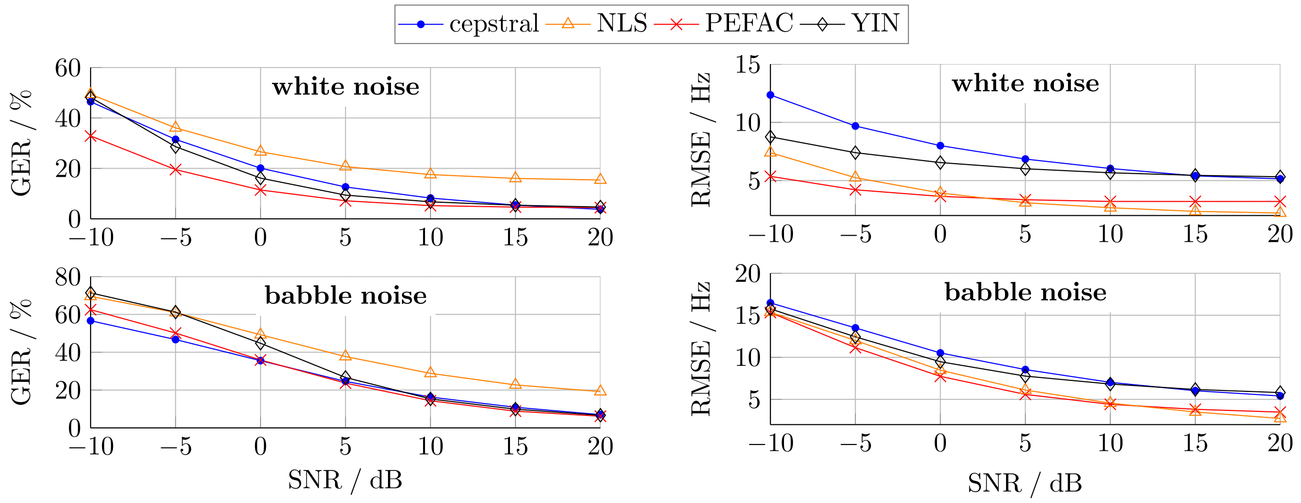
$$G(\hat{f}_0[\ell]) = \begin{cases} 1, & \text{if } |\hat{f}_0[\ell] - f_0[\ell]|/f_0[\ell] \geq \Delta \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The threshold Δ is set to 20 % in our experiments. The RMSE is defined as standard deviation between the estimated and the annotated frequency and is only evaluated on voiced blocks where no gross error occurred, i. e. $G(\hat{f}_0[\ell]) = 0$. The computational complexity is evaluated in Matlab using time measurements on a state-of-the-art desktop PC. In this contribution, we present the raw performance of the estimators. Thus, post-processing steps like dynamic programming are deactivated.

Results

Figure 1(a) shows the GER in white and babble noise for different SNRs in a non-reverberant environment. Except for the NLS algorithm, which obtains considerably higher GERs, all candidates show a similar performance in high SNR conditions. Considering white noise at low SNRs, PEFAC obtains the lowest error rates. In babble noise, however, the performance of the cepstral peak picking method is similar to PEFAC.

This work was funded by the DFG Research Grant GE2538/2-1.



(a) GER

(b) RMSE

Figure 1: GER and RMSE for white and babble noise depending on the SNR without reverberation.

Figure 1(b) shows the corresponding results for the RMSE. In white noise, PEFAC and NLS obtain considerably lower error rates than the cepstral peak picking algorithm and YIN. This point can also be made for babble noise at high SNRs. In babble noise and at low SNRs the RMSE performance of all candidates is similar and worse than for white noise.

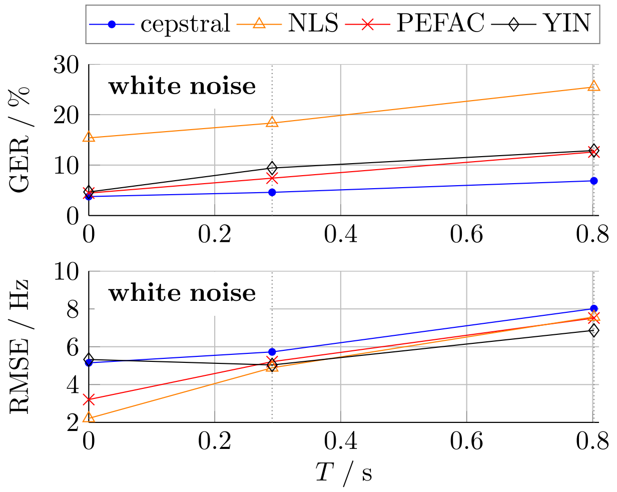


Figure 2: GER and RMSE for white noise at 20 dB SNR depending on the reverberation time.

Figure 2 shows the GER and RMSE depending on the reverberation time T in white noise at 20 dB SNR. Here, we omit the result in babble noise as the observable effects are quite similar for both noise types. The reverberation decreases the performance of the fundamental frequency estimators which leads to an increase of 5 % to 10 % of the GER. One exception is the cepstral peak picking algorithm which appears to be more robust towards reverberation than the other candidates.

Table 1: Processing time of fundamental frequency estimators.

	cepstral	NLS	PEFAC	YIN
time / min	0:17	15:38	3:15	1:17

Finally, Table 1 shows the results of the time measurements in MATLAB which serve as an assessment of the computational complexity. For this, a database with a duration 19:40 min was used. The results indicate that the NLS approach exhibits the largest complexity, while the cepstral approach is most efficient.

Conclusions

In this paper, we presented a comparison of state-of-the-art fundamental frequency estimators. In non-reverberant environments, the lowest fine pitch and gross error rates have been obtained for PEFAC. The cepstral peak picking algorithm appears to be most robust towards reverberation. Furthermore, the cepstral peak picking method is computationally most efficient, while the complexity of NLS is the largest.

References

[1] M. Krawczyk and T. Gerkmann. “STFT Phase Improvement for Single Channel Speech Enhancement”. In: *International Workshop on Acoustic Signal Enhancement, 2012.* Aachen, Germany, Sept. 2012.

[2] A. M. Noll. “Cepstrum Pitch Determination”. In: *The Journal of the Acoustical Society of America* 44.6 (1968), pp. 1585–1591.

[3] S. Gonzalez and M. Brookes. “PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.2 (Feb. 2014), pp. 518–530.

[4] A. d. Cheveigné and H. Kawahara. “YIN, a fundamental frequency estimator for speech and music”. In: *The Journal of the Acoustical Society of America* 111.4 (2002), pp. 1917–1930.

[5] J. Tabrikian, S. Dubnov, and Y. Dickalov. “Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model”. In: *IEEE Transactions on Speech and Audio Processing* 12.1 (2004), pp. 76–87.

[6] M. G. Christensen and A. Jakobsson. *Multi-Pitch Estimation.* Vol. 5. Synthesis Lectures on Speech & Audio Processing. Morgan and Claypool Publishers, 2009.

[7] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. *TIMIT Acoustic-Phonetic Continuous Speech Corpus.* Linguistic Data Consortium, Philadelphia, 1993.

[8] S. Gonzalez and M. Brookes. *Pitch of the Core TIMIT database set.* 2011. URL: `http://www.ee.ic.ac.uk/hp/staff/dmb/data/TIMITfxv.zip`.

[9] H. J. M. Steeneken and F. W. M. Geurtsen. *Description of the RSG.10 noise database.* Technical Report IZF 1988-3. TNO Institute for perception, 1988.

[10] M. Jeub, M. Schafer, and P. Vary. “A binaural room impulse response database for the evaluation of dereverberation algorithms”. In: *16th International Conference on Digital Signal Processing, 2009.* Santorini, Greece, July 2009.