Improving continuous F0 estimator with adaptive time-warping for high-quality speech synthesis

Mohammed Salah Al-Radhi[1], *Tamás Gábor Csapó*[1,2], *Géza Németh*[1]

[1]Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary
[2]MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary
{malradhi,csapot,nemeth}@tmit.bme.hu

Abstract

Parametric representation of speech often implies fundamental frequency (also referred to F0 or pitch) contour as a part of the text-to-speech (TTS) synthesis. During voiced speech such as vowels, pitch values can be successfully estimated over a short-time period (e.g., a speech frame of 25ms). But in unvoiced speech such as unvoiced consonants, the long term spectrum of turbulent airflow tends to be a weak function of frequency [1], which suggests that the identification of a single reliable F0 value in unvoiced regions is not possible.

In recent years, there has been a rising trend of assuming that continuous F0 observations are present similarly in unvoiced regions and there have been various modelling schemes along these lines. Garner et al. [2], the baseline method in this study, proposed a simple continuous F0 (ContF0) tracker, where the measurement distribution is determined from the autocorrelation coefficients. Tóth and Csapó [3] have shown that the ContF0 contour can be predicted better with hidden Markov model (HMM) and feed-forward deep neural network (FF-DNN) methods than traditional discontinuous F0, in case of statistical parametric speech synthesis. However, ContF0 contour is still sensitive to additive noise in speech signals and suffer from short-term errors (when the ContF0 changes rather quickly over time).

In this work, we alleviate these problems by employing adaptive time-warping method [4]. We iteratively apply a time axis warping on the input signal with weighted averaging process. Two English speakers were chosen from the CMU-ARCTIC database [5], denoted BDL (American English, male), and SLT (American English, female), each one consisting of 1132 sentences; one channel was the waveform, the other laryngograph (from which a reliable pitch estimate can be derived). 20 sentences from each speaker were chosen randomly to be analyzed and synthesized with the baseline and proposed method. In the evaluation, the ground truth is extracted from electro-glottal graph (EGG) signals using Praat. An example of ContF0 estimation on a female speech is shown in Figure 1. The trajectory given by the proposed method is in general smoother and as the example in Figure 1 shows, is not influenced by the dip at frame 32 and frame 138. Moreover, we used white Gaussian noise as the additive background noise to test the quality of the adContF0 (adaptive continuous) method and also to clarify the effects of refinement. The amount of noise is specified by signal-to-noise ratio (SNR) ranged from -40 to 30 dB. We calculated the normalized root mean square error (NRMSE) over selected sentences for each speaker. It can be clearly seen from Figure 2 that the NRMSE values for the proposed method are smaller and outperformed the conventional one, as we expected. Hence, the improved adContF0 pitch estimation algorithm achieves higher accuracy on noisy and clean speech than the baseline.

The results of this study can potentially be useful for a number of speech technologies, including statistical parametric speech synthesis and voice conversion.

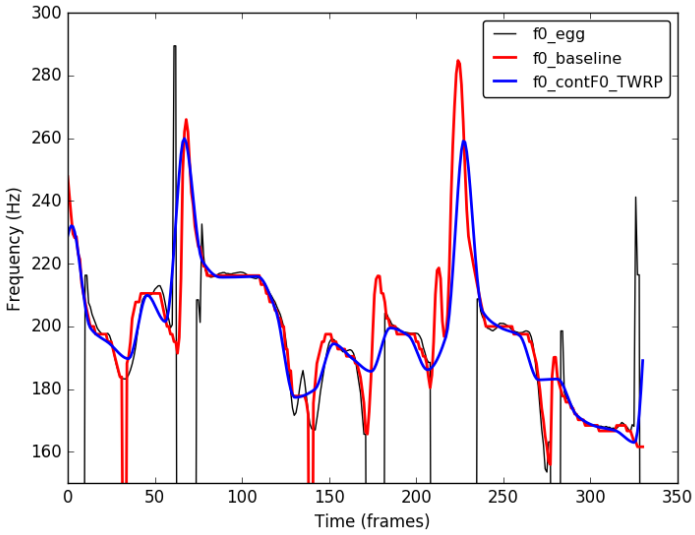

Figure 1: F0 trajectories estimated by the ground truth (black), baseline (red), and plotted along with proposed method (blue). Sentence: “Everything was working smoothly, better than I had expected.”, from speaker SLT.

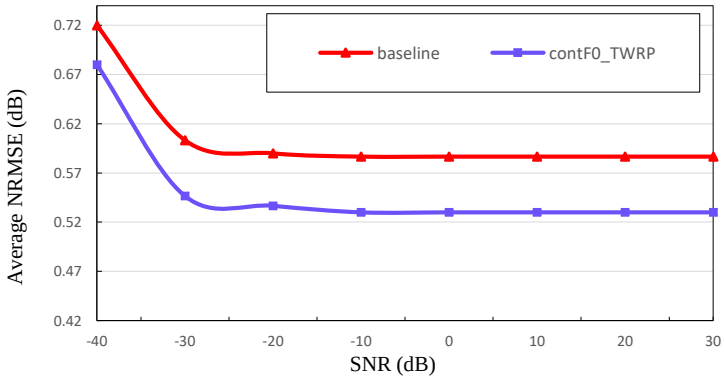

Figure 2: Average NRMSE of refined ContF0 estimation vs. additive noise SNR. The baseline estimate (red) error deviations were reduced by a factor of 5.4% (blue) with time-warping method.

References

[1] Talkin D., "A robust algorithm for pitch tracking (RAPT), "Speech coding and synthesis, Elsevier, p. 495–518, 1995.

[2] Garner P. N., Cernak M., and Motlicek P., "A simple continuous pitch estimation algorithm," IEEE Signal Processing Letters, vol. 20, no. 1, pp. 102-105, 2013.

[3] Bálint Pál Tóth and Tamás Gábor Csapó, "Continuous Fundamental Frequency Prediction with Deep Neural Networks," in Proc. of the European Signal Processing Conference (EUSIPCO), Budapest, Hungary, pp. 1348-1352, 2016.

[4] Kawahara H., Agiomyrgiannakis Y., and Zen H., "Using instantaneous frequency and aperiodicity detection to estimate f0 for high-quality speech synthesis," in 9th ISCA Workshop on Speech Synthesis, CA, USA, 2016

[5] Kominek J. and Black A. W., "CMU ARCTIC databases for speech synthesis," Carnegie Mellon University, 2003.