Multi-Microphone Periodicity Function for Robust F0 Estimation in Real Noisy and Reverberant Environments

Federico Flego, Maurizio Omologo

ITC-irst (Centro per la Ricerca Scientifica e Tecnologica) I-38050 Povo - Trento (Italy)

[flego,omologo]@itc.it

Abstract

This paper outlines a new method to extract F0 from distant-talking speech signals acquired by a microphone network, which exploits the redundancy across the signals proceeding from each microphone, by jointly processing the different contributes. To this purpose, a multi-microphone periodicity function is derived from the magnitude spectrum computed on each microphone signal. This function allows to estimate F0 reliably, even under reverberant conditions, without the need of any post-processing or smoothing technique. Experiments, conducted on real lectures, showed that the proposed frequency-domain algorithm is more suitable than other time-domain based ones.

Index Terms: speech analysis, fundamental frequency estimation, multi-microphone processing, distant-talking interaction.

1. Introduction

In the CHIL project, various signal processing techniques are being investigated that aim to address challenging problems among which acoustic event classification, speaker localization and tracking, distant-talking speech recognition, speech activity detection, speaker identification and verification [1].

One way to pursue all these objectives is that of deriving a model of the source (e.g. the speaker) from the given multi-microphone data. To this purpose, a Distributed Microphone Network (DMN) is used, which consists in a generic set of microphones localized in space without any specific geometry.

In this work we address the problem of deriving a robust estimation of the fundamental frequency F0 from the variety of signals recorded through the microphone network. Speech signals recorded by microphones placed far from a talker are severely degraded by both background noise and reverberation, which depends on spatial relationships among the microphones and the talker, as well as on the scenario acoustic characteristics.

Estimating F0 independently for each microphone signal and applying then majority vote or other fusion based methods may represent a possible approach. Another way to perform F0 estimation is to extend to the multi-microphone case a paradigm that works for a single microphone close-talking case. A time-domain F0 extraction algorithm based on Weighted Autocorrelation (WAUTOC) [2] was experimented in the past [3], which showed good performance on a real multi-microphone database of distant-talking speech sequences reproduced in an office environment. In particular, the resulting multi-microphone WAUTOC technique offers the advantage of obtaining better performance than single microphone based processing, without any assumption or knowledge about the position of the microphones as well as of the talker. However, a deep analysis of the results showed that the given time-domain solution was still penalized by reverberation effects which introduce phenomena difficult to model and to circumvent by working in the time-domain. Hence a frequency domain approach was investigated to better exploit the fine pitch structure that is common to the given microphone signals. In this work, an algorithm based on a Multi-microphone Periodicity Function (MPF) is then introduced and compared to the multi-microphone WAUTOC and to a multi-microphone extension of the YIN algorithm [4].

Experiments were conducted on a real corpus of lectures recorded in a noisy and reverberant environment, which was used in 2005 at NIST for benchmarking purposes (for further details see http://www.nist.gov/speech/tests/rt/rt2005/spring). Results show the advantages of the proposed MPF algorithm.

The paper is organized as follows: Section 2 introduces the MPF based F0 extraction algorithm; Section 3 and 4 present the multi-microphone YIN and WAUTOC algorithms, respectively; Section 5 and 6 describe the given experimental set-up and the evaluation criteria; Section 7 reports on the experimental results that were obtained and Section 8 draws some conclusions and outlines future work.

2. MPF based F0 extraction

The F0 extraction algorithm here outlined can be classified under the frequency-domain category and, in particular, it includes a processing that resembles that described in [5].

Given the above mentioned DMN context, the different paths, from the source to each microphone, are affected differently by the non linear reverberation effects, which can enhance some frequencies while attenuating others. The peaks in the magnitude spectrum which refer to F0 and its harmonics, are thus altered in dynamics but preserved in frequency location. Hence, the common harmonic structure across the different magnitude spectra, can be exploited for better estimating the fundamental frequency.

Let $x_i(n)$ be the downsampled version of the source speech signal recorded at the i-th microphone of M microphones and $w(n)$ a window function of length L_w samples. For each analysis frame, the windowed signal is zero-padded to produce the vector X_i^w of length L_f. An FFT is then computed and its absolute value is derived as follows:

$$S_i(f_k) = \left|\mathrm{FFT}\{X_i^w\}(k)\right|^\gamma, \quad 1 \le k \le L_f. \tag{1}$$

being f_k the k-th frequency bin and γ a spectral compression factor ($\gamma = 2$ returns the power spectrum). Next step is to compute a

This work was partially funded by the EU under the Integrated Project CHIL (IP 506909). http://chil.server.de

sum of the real valued normalized functions $S_i(f_k)$:

$$\overline{S}(f_k) = \sum_{i=1}^{M} \frac{S_i(f_k)}{\max_k\{S_i(f_k)\}}, \quad 1 \leq k \leq \frac{L_f}{2} + 1. \qquad (2)$$

Next, IFFT is applied to obtain the *Multi-microphone Periodicity Function* $\overline{s}(\tau)$ in the lag-domain

$$\overline{s}(\tau) = \mathrm{IFFT}\{\overline{S}([f_1, \ldots, f_{L_f/2+1}, f_{L_f/2}, \ldots, f_2])\}, \qquad (3)$$

where the argument of the IFFT is a vector whose L_f elements are the $\overline{S}(f_k)$ values, with k first ranging from 1 to $L_f/2+1$, then decreasing from $L_f/2$ to 2, so that the original symmetry of $S_i(f_k)$ is restored. Resulting thus $\overline{s}(\tau)$ a minimum phase signal, the lag value at which a maximum is found can be considered the fundamental frequency period T_0 estimated for the analysed frame. After applying interpolation to improve lag resolution, $\overline{s'}(\tau)$ is obtained and it holds that

$$T_0 = \arg\max_{\tau}\{\overline{s'}(\tau)\}, \quad T_{\min} \leq \tau \leq T_{\max}, \qquad (4)$$

where $T_{\min}$ and $T_{\max}$ are the minimum and maximum fundamental frequency period.

3. A multi-microphone version of YIN

The *YIN* algorithm is a state-of-the-art time-domain based algorithm derived from the autocorrelation function and was designed to work on a single microphone signal.

As described in [4], first the difference function, $d_i(\tau)$, is derived

$$d_i(\tau) = \sum_n [x_i(n) - x_i(n+\tau)]^2, \qquad (5)$$

being n the time index in the analysis frame and i the microphone index. This function is less sensitive to changes in signal amplitudes, compared to the autocorrelation function, thus being less prone to "too low/too high" F0 estimation errors. In addition, in order to further reduce errors, the *cumulative mean normalized difference function* is computed

$$d'_i(\tau) = \begin{cases} 1, & \text{if } \tau = 0, \\ d_i(\tau)/[(1/\tau)\sum_{j=1}^{\tau} d_i(j)], & \text{otherwise}, \end{cases} \qquad (6)$$

and a higher performance is thus reported.

A *YIN* multi-microphone version is derived here by simply normalizing the difference function computed for each microphone signal, $d_i(\tau)$, and then by averaging over all microphones

$$d_M(\tau) = \frac{1}{M}\sum_{i=1}^{M} \frac{d_i(\tau)}{\max_{\tau}\{d_i(\tau)\}} \qquad (7)$$

The *cumulative mean normalized difference function* turns then into

$$d''(\tau) = \begin{cases} 1, & \text{if } \tau = 0, \\ d_M(\tau)/[(1/\tau)\sum_{j=1}^{\tau} d_M(j)], & \text{otherwise}, \end{cases} \qquad (8)$$

which is then used instead of (6).

Although other alternatives had been explored (e.g., by averaging the cumulative mean normalized difference function) preliminary experiments showed that the approach based on equation (8) gave the best performance.

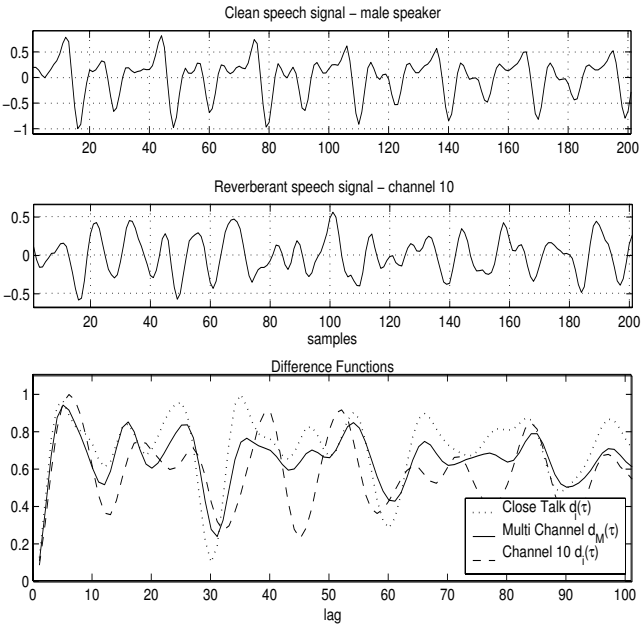

Figure 1: *Top: example of a vowel portion extracted from a close-talk recording. Middle: same speech segment captured from a distant microphone. Bottom: Difference function $d_i(\tau)$ computed on the close-talk and on the far microphone signal, and multi-microphone difference function $d_M(\tau)$ computed on the whole set of microphones.*

Please let us note that the proposed extension of YIN to the multi-microphone case does not represent a ultimate best YIN-based solution to the given problem. For instance, a specific work, outside the scope of this paper, should be conducted to check if a more effective postprocessing can be conceived in this case. In order to show the plausibility of the outlined choice (equation 8), Figure 1 shows an example that justifies the here investigated multi-microphone version of YIN . A considerable mismatch can be observed in the time-domain structure of the close-talk and of the reverberated far-microphone sequence for the given frame (the microphone was at 3 meter distance from the speaker). Then, the figure reports on the comparison between the difference functions obtained by applying YIN to the close-talk and to the far microphone signals and by applying the multi-microphone YIN algorithm to the entire set of 16 far microphone signals. One can note that the minimum, located at 30 samples and clearly missed when processing the far microphone signal (a value between 40 and 50 was chosen), is eventually recovered thanks to the effectiveness of the multi-microphone YIN processing.

4. WAUTOC-based F0 estimation

In the past, many F0 (or pitch) estimation methods were proposed and evaluated [4, 6]. Some of these methods derive from the basic formulations of short-term autocorrelation and AMDF functions. The weighted autocorrelation based one, recently introduced in [2], proved to be particularly robust to noise and also to doubling or halving period estimation mismatch. The WAUTOC function is defined as:

$$wautoc_i(\tau) = \frac{\sum_{n=0}^{N-\tau-1} x_i(n)x_i(n+\tau)}{\sum_{n=0}^{N-\tau-1} |x_i(n) - x_i(n+\tau)| + \epsilon}, \qquad (9)$$

being i the microphone index and ϵ a constant value that prevents the function from getting too high dynamics or zero-division condition. In practice, the denominator of the fraction in (9) corresponds to an AMDF function while the numerator represents an autocorrelation function.

Since in correspondence of the pitch period the autocorrelation and the AMDF functions have, respectively, a maximum and a minimum, WAUTOC-based F0 estimation takes benefits from the characteristics of both functions. The pitch period is estimated as

$$l = \arg\max_{\tau}\{wautoc_i(\tau)\}. \quad (10)$$

As introduced in [3], to extend the WAUTOC-based technique to the multi-microphone case, a suitable way is that of averaging the given function over the entire microphone network, which leads to the computation of

$$f(\tau) = \sum_{i=1}^{M} wautoc_i(\tau), \quad (11)$$

where M denotes the microphone number.

5. Experimental set-up

The proposed algorithm was tested on the spontaneous speech corpora collected under the CHIL project[1], which consist of 13 recordings, each about 5 minutes long, from female and male speakers extracted from real seminar sessions held at the Karlsruhe University.

Each speaker, wore a "Countryman E6" close-talking microphone, to capture a noise-free, non reverberant speech signal, and moved freely in the area labeled "speaker area", showed in Figure 2. The DMN layout, as shown in the figure, employs four inverted "T"-shaped microphone arrays, each consisting of 4 microphones. Inter-microphone spacing is 20 cm and 30 cm along the horizontal and vertical directions, respectively. Speech sequences were recorded with $44.1\ kHz$ sampling rate and 16 bit resolution.

The room is $7.10\ m\ \times\ 5.90\ m$ wide and the ceiling height is $3\ m$. There is one entrance in the north wall, and two more doors in the south wall leading to other offices. Reverberation time was $T_{60} \simeq 0.45s$.

During recordings there were audience seated on the chairs placed as depicted in the figure.

6. F0 evaluation criteria

To evaluate the proposed algorithm, reliable "ground-truth" pitch estimates were derived. This was accomplished applying three existing pitch extractor algorithms, Praat, SFS and WaveSurfer [7, 8, 9] to the the close-talk recordings, obtaining a set of three estimates for each $10\ ms$ length frame. Then, only the frames for which the variance of the three estimates was $< 3\ Hz$ were marked as "voiced", associating to each of them the F0 obtained averaging the three reference values.

A frequently used method to compare the performance between different algorithms is to compute the Gross Error Rate (GER). This is calculated considering the number of F0 estimates which differ by more than a certain percentage from the reference values.

[1]A description of the used speech corpora can be found at http://chil.server.de, http://www.nist.gov/speech and http://www.clear-evaluation.org.

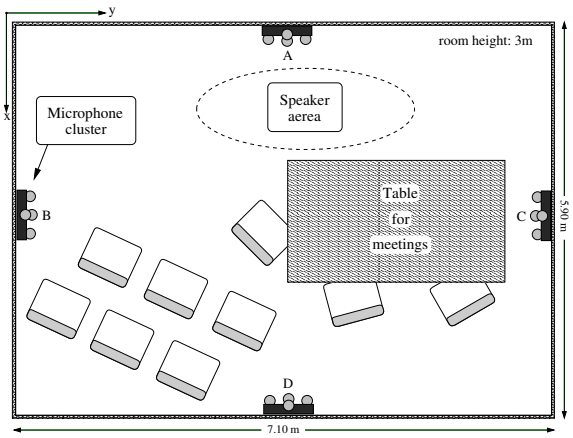

Figure 2: *The CHIL seminar and meeting room at the Karlsruhe University where the four inverted "T"-shaped microphone arrays are labeled with letters "A", "B", "C" and "D".*

In this work thresholds of 20% and 5% are used for the GER estimation, indicated with GER(20) and GER(5), respectively. The reason for this choice is that, if a pitch estimate satisfies this criteria, then several techniques can be used to refine its value, as it was pointed out in [4].

7. Experimental results

In the following experiments, F0 estimates were obtained using an analysis step of $10\ ms$ and an analysis window of $40\ ms$ length for the YIN algorithm and of $60\ ms$ for the WAUTOC and MPF (Hamming window) algorithms. These different window durations were determined by preliminary experiments aimed to optimize each algorithm performance.

A first experiment was conducted to assess the effect of spectral compression on the GER, obtained varying γ in Equation 1. As shown in Figure 3, when the MPF algorithm is applied to the close-talk signals, the lowest GER(20) is obtained for $\gamma = 1.7$. However, considering the resulting y-axis range and the curve flatness, the test also points out the weak dependency of GER(20) on the γ parameter. In the case of reverberant signals instead, the lower panel in the figure indicates that $\gamma = 0.5$ provides the best GER(20), and that the latter increases rapidly for small variations of γ. These results are in accordance with those reported in [10], where spectral compression with $\gamma < 1$ was proved to be beneficial for pitch estimation applied to both speech and musical signals.

After $\gamma = 0.5$ was set in the MPF algorithm, the three algorithms were tested in their single-microphone version (dotted line) as well as in their multi-microphone version (continuous line) and the results are reported in Figure 4.

From the results, it can be observed that YIN performed best when applied to the close-talk speech signal (dashed line). However, in the multi-microphone and single-microphone case, MPF performed better than the two other algorithms.

Moreover, applying to far microphone signals any of the algorithms in single-microphone fashion always led to a performance worse than that obtained using the MPF based algorithm. The above considerations hold for both the GER(20) measure (upper panel) and the GER(5) measure (lower panel), although in the latter case the superiority of MPF over the multi-microphone YIN is

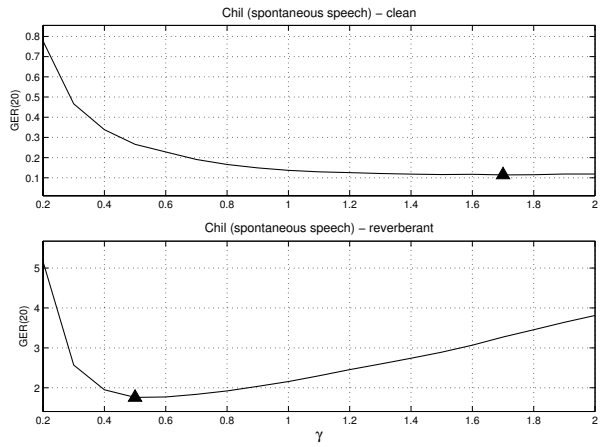

Figure 3: *Dependency of the MPF GER(20) on the parameter γ of Equation (1)*

less evident.

8. Conclusions and Future Work

This paper addressed the problem of estimating the fundamental frequency on distant-talking speech, given a set of microphones distributed in space.

Although signals are degraded by noise and reverberation, it is shown that the use of the proposed MPF multi-microphone algorithm allows to obtain a remarkable reduction in gross error rates, which represents a promising starting point for future research activities.

It is also worth noting that applying the MPF-based algorithm blindly is straightforward; on the other hand, applying a single microphone algorithm to every far microphone signal would in any case require a further processing to select the most reliable F0 among the resulting candidates.

Next steps include the objective to exploit the resulting F0 estimates provided by the MPF function as features for acoustic event detection and classification, speech activity detection, and eventually distant-talking ASR.

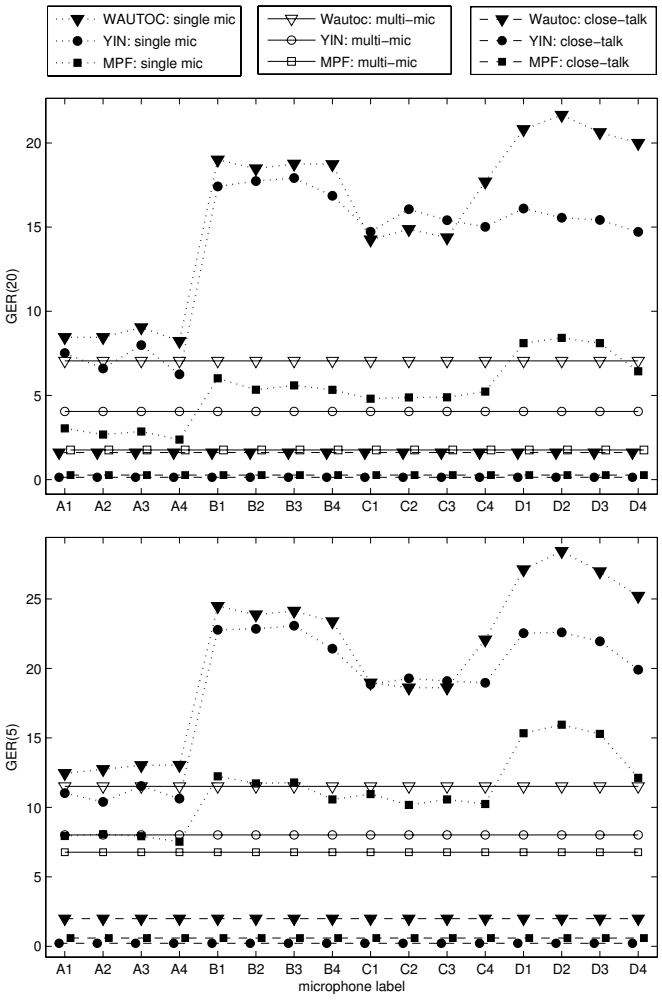

Figure 4: *The three curves show the gross error rates derived by applying WAUTOC, YIN and MPF, respectively, to each of the 16 microphone signals. The six horizontal lines indicate the performance provided by each of the three algorithms on close-talking signals (dashed lines) and the corresponding performance obtained by their multi-microphone version applied to all the far microphone signals (continuous line). GER(20) and GER(5) are reported in the upper and lower panel, respectively.*

9. References

[1] D. Macho et al., "Automatic Speech Activity Detection, Source Localization, and Speech Recognition on the CHIL Seminar Corpus", *ICME* Conference, 2005.

[2] T. Shimamura, H. Kobayashi, "Weighted Autocorrelation for Pitch Extraction of Noisy Speech", IEEE Trans. on Speech and Audio Processing, vol. 9, n. 7, pp. 727–730, October 2001.

[3] F. Flego, M. Omologo, L. Armani, "On the Use of a Weighted Autocorrelation Based Fundamental Frequency Estimation for a Multidimensional Speech Input", *Proc. of ICSLP*, 2004.

[4] A. de Cheveigné, H. Kawahara, "YIN, a Fundamental Frequency Estimator for Speech and Music", J. Acoust. Soc. Am., Apr. 2002.

[5] S. Sagayama, S. Furui, "Pitch Extraction Using the Lag Window Method", *Proc. of IECEJ* Meeting, 1978 (in Japanese).

[6] W. Hess, "Pitch Determination of Speech Signals", Springer, 1983.

[7] P. Boersma, D. Weenink, "Praat: doing phonetics by computer (Version 4.4.16, 2006)" [Computer program]. Retrieved from http://www.praat.org

[8] M. Huckvale, "SFS: Speech Filing System" [Computer program]. Retrieved from http://www.phon.ucl.ac.uk/resource/sfs.html

[9] K. Sjölander, J. Beskow, "WaveSurfer - an open source speech tool", *Proc. of ICSLP*, 2000.

[10] T. Tolonen, M. Karjalainen, "A Computationally Efficient Multipitch Analysis Model", *IEEE trans. Speech and Audio Processing*, 2000.