14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

ROBUST F0 ESTIMATION BASED ON A MULTI-MICROPHONE PERIODICITY FUNCTION FOR DISTANT-TALKING SPEECH

Federico Flego, Maurizio Omologo

ITC-irst (Centro per la Ricerca Scientifica e Tecnologica) I-38050 Povo - Trento (Italy)
[flego,omologo]@itc.it

ABSTRACT

This work addresses the problem of deriving F0 from distant-talking speech signals acquired by a microphone network. The method here proposed exploits the redundancy across the channels by jointly processing the different signals. To this purpose, a multi-microphone periodicity function is derived from the magnitude spectrum of all the channels. 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 data, showed that the proposed frequency-domain algorithm is more suitable than other time-domain based ones.

1. INTRODUCTION

An attractive future scenario consists in the development of new workspaces where the so-called "ambient intelligence" is realized through a wide usage of sensors (cameras, microphones, etc.) connected to computers that fade in the background, largely invisible and significantly less intrusive to humans.

In the CHIL project, a Distributed Microphone Network is used, which consists in a generic set of microphones localized in space without any specific geometry. The analysis of the resulting acoustic scenario is accomplished by a multi-channel processing aimed at extracting real-time information for speaker tracking, acoustic event classification, and distant-talking speech recognition [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. In particular, 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.

Estimating F0 independently for each channel 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, 4], 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 [5]. Experimental results show the advantages of the proposed 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 [6].

Given the above mentioned Distributed Microphone Network 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:

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

being f_k the frequency bin with index k. Next step is to compute a weighted sum of the real valued normalized functions $S_i(f_k)$:

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

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

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

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

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

$L_f/2+1$, 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

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

where $T_{\min}$ and $T_{\max}$ are the minimum and maximum fundamental frequency period. To assign the weight values, c_i, first a reference spectrum $S_P(f_k)$ is estimated as a product of channel magnitude spectrum in the following way

$$S_P(f_k) = \prod_{i=1}^{M} \frac{S_i(f_k)}{\max_k\{S_i(f_k)\}}, \tag{5}$$

and then each weight c_i is derived basing on the Cauchy-Schwarz inequality applied to $S_i(f_k)$ and $S_P(f_k)$ considering them as if they were vectors:

$$c_i = \frac{\sum_{k=1}^{K} S_P(f_k)S_i(f_k)}{\sqrt{\sum_{k=1}^{K} S_P^2(f_k)}\sqrt{\sum_{k=1}^{K} S_i^2(f_k)}}, \quad K = \frac{L_f}{2}+1. \tag{6}$$

In this way, $S_P(f_k)$ will retain information common to the different channels while rejecting interference showing different frequency patterns not common to all channels.

Coefficients c_i will thus range from 0 to 1, and have been introduced to represent the reliability of each channel spectrum $S_i(f_k)$, which may depend on the speaker position, head orientation or on the presence of other sources of noise. As discussed in Section 5, white noise sequences at different SNR were added to some given recordings, to confirm the usefulness of coefficients c_i in the case when a few microphones are affected by a lower SNR.

3. A MULTI-MICROPHONE VERSION OF YIN

The *YIN* algorithm is a time-domain based algorithm derived from the autocorrelation function and was designed to work on a single channel source. This algorithm represents one of the state of the art pitch detection algorithms and was thus chosen here for comparison purposes.

As described in [5], first the difference function, $d_i(\tau)$, is derived

$$d_i(\tau) = \sum_n [x_i(n) - x_i(n+\tau)]^2, \tag{7}$$

being n the time index in the analysis frame and i the microphone index. The author demonstrates that 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 derived

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

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 contributes

$$d_M(\tau) = \frac{1}{M}\sum_{i=1}^{M} \frac{d_i(\tau)}{\max_{\tau}\{d_i(\tau)\}} \tag{9}$$

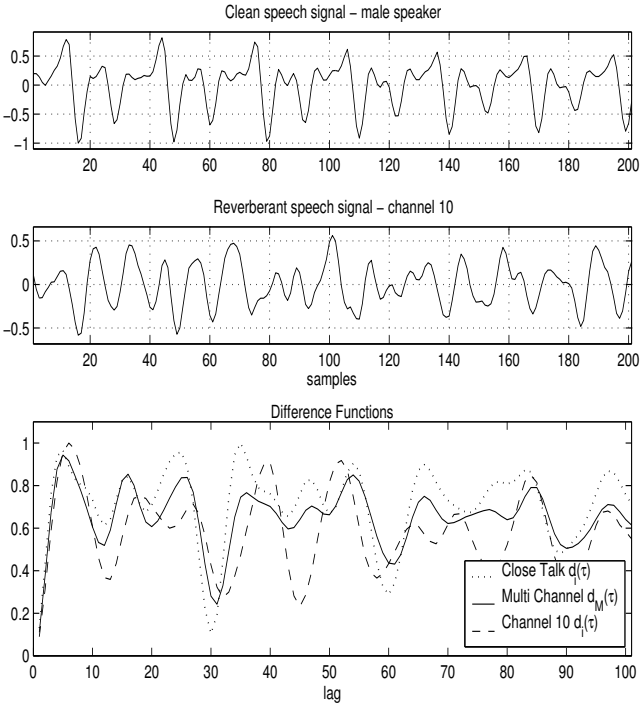

Figure 1: *Top: example of a vowel portion extracted from a close-talk recording. Middle: same speech segment captured from a distant microphone. Bottom: Difference function $d_i(\tau)$ computed on the close-talk and on the far microphone signal, and multi-microphone difference function $d_M(\tau)$ computed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

The *cumulative mean normalized difference function* turns then into

$$d''(\tau) = \begin{cases} 1, & \text{if } \tau = 0, \\ d_M(\tau)/[(1/\tau)\sum_{j=1}^{\tau} d_M(j)], & \text{otherwise}, \end{cases} \tag{10}$$

which is then used instead of (8).

Other alternatives had been explored, as for instance averaging the cumulative mean normalized difference function rather than the difference function. In some preliminary experiments the approach based on equation (10) gave the best performance.

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 (see details on the various steps of the algorithm in [5]). However, in order to show the plausibility of the outlined choice (equation 10), 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 10 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 [7, 5, 8]. Some of these methods derive from the basic formulations of short-term autocorrelation and AMDF functions.

The weighted autocorrelation-based one, recently introduced by [2], proved to be particulary robust to noise and also to doubling or halving period estimation mismatch. The WAUTOC function is defined as:

$$wautoc_i(\tau) = \frac{\sum_{n=0}^{N-\tau-1} x_i(n)x_i(n+\tau)}{\sum_{n=0}^{N-\tau-1} |x_i(n) - x_i(n+\tau)| + \epsilon}, \qquad (11)$$

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 (11) 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

$$l = \arg\max_{\tau}\{wautoc_i(\tau)\}. \qquad (12)$$

As introduced in [4], 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

$$f(\tau) = \sum_{i=1}^{M} wautoc_i(\tau), \qquad (13)$$

where M denotes the number of microphone channels.

5. EXPERIMENTAL SET-UP

In order to measure the performance of the proposed algorithm, the Keele database was used [9], which consists of five male and five female English speakers who pronounced phonetically balanced sentences. The total duration of the database is 9 minutes.

Since a multi-microphone database was needed, the Keele database was reproduced by using a very high quality dual-concentric (TANNOY 600A) loudspeaker, placed in two positions ($P1$ and $P2$) in order to have different sound propagation situations. Speech sequences were then recorded using 10 omnidirectional microphones and a multi-microphone platform operating at 20 kHz and 16 bit[1].

The office is 3 m x 7 m wide and 3 m high and is characterized by a reverberation time $T_{60} \simeq 0.35s$. As shown in Figure 2, adjacent microphones were from 0.2 to 2 meters far each other. During recordings there were no people in the room, and the only source of noise was the computer fan.

6. F0 EVALUATION CRITERIA

An F0 estimation algorithm can be evaluated in different ways according to the application purposes. One of the most common ways is that of using a laringograph as reference from which a reliable "ground-truth" estimate can be derived [5]. Generally, the reference F0 is extracted automatically from the laringograph output and then manually checked (visually) in order to avoid discrepancies in irregular voiced portions.

[1]This multi-microphone database can be downloaded at: http://shine.itc.it

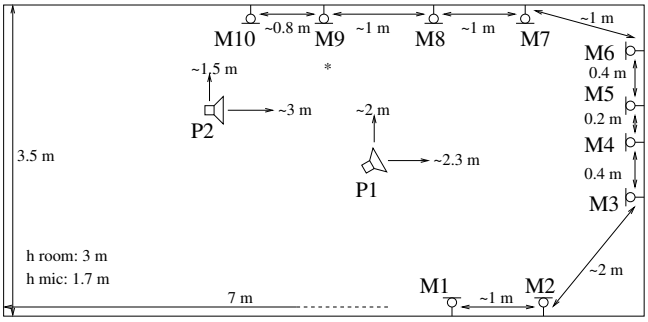

Figure 2: *The office with ten microphones and the loudspeaker placed in two positions, one left to right and the other 30 degrees top right. The room is quiet, except for a computer fan marked with a *.*

To be more accurate in our tests, an analysis step of $1ms$ was chosen and, given that the original labels were derived every $10ms$, a new reference F0 was derived using the laryngograph signals included in the Keele database. This was accomplished by pre-filtering these signals with a high-pass filter to eliminate a slowly varying bias present, probably due to movement of the speaker during original recordings. Then the Praat program [10] was used to obtain the new references and the result was manually checked to correct errors.

F0 candidates are evaluated only for analysis frames manually labeled as "voiced".

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 laringograph reference values. In this work a threshold of 20% is used for the GER estimation. The reason for this choice is that, if a pitch estimate satisfies this criteria, then several techniques can be used to refine its value, [5].

7. EXPERIMENTAL RESULTS

In the following experiments, F0 estimates were obtained using an analysis step of 1 ms and an analysis window of 30 ms length for the WAUTOC and the YIN algorithms and of 60 ms length (Hamming window) for the MPF based algorithm. These different window durations were determined by preliminary experiments aimed to optimize each algorithm performance.

A first experiment was conducted to assess the effectiveness of introducing the weights c_i in (2) and results are showed in Figure 3. Microphones 1, 2 and 3 were chosen to form a subset, from which two other database replica were derived adding white noise with SNR of 0 and -5 dB to one of the microphone signals.

The multi-microphone WAUTOC and YIN as well as the MPF based algorithm were tested under those three conditions. In particular MPF was run twice, first with all weights c_i set to 1, then applying equation (6). The results showed the effectiveness of the latter setting and, in practice, the capability of the proposed algorithm to exploit input channels with a better signal quality.

Figure 3 shows the gross error rate provided by the three algorithms in their single-microphone version as well as in their multi-microphone version. The whole set of microphone signals were considered. From those results, one can observe that YIN is the best algorithm when applied to the close-talk speech signal. However, in the multi-microphone

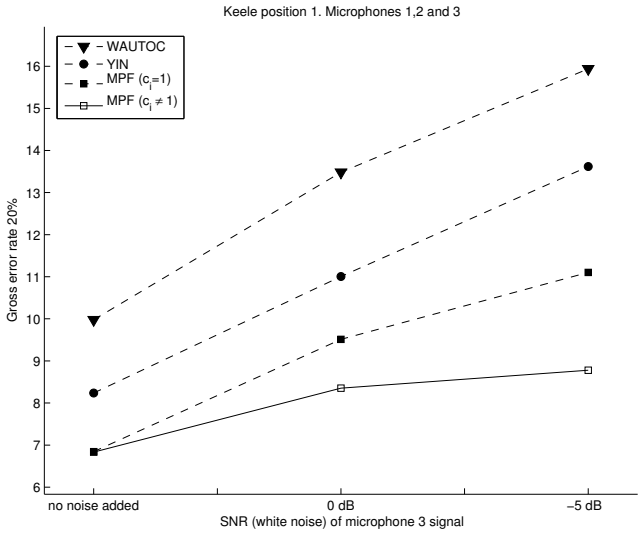

Figure 3: *Gross error rates obtained by the multi-microphone version of each algorithm under different noisy conditions. Only three microphones were used and white noise was added to channel 3 at different SNRs.*

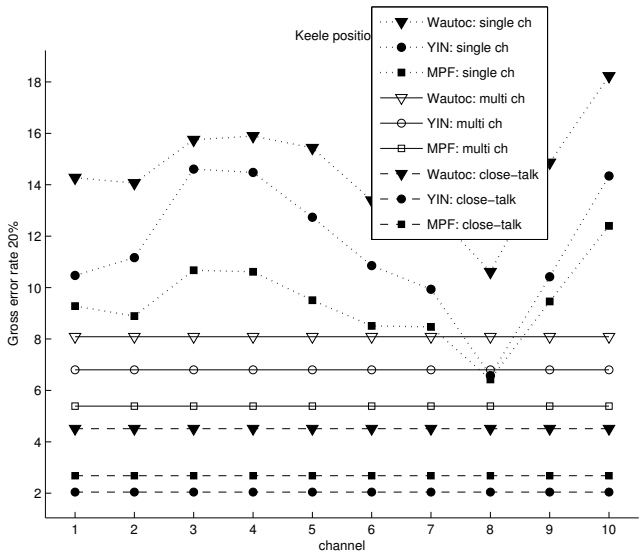

Figure 4: *The three curves show the gross error rates derived by applying WAUTOC, YIN and MPF, respectively, to each of the 10 microphone signals. Results refer to the loudspeaker position P1. The six horizontal lines indicate the performance provided by each of the three algorithms on close-talking signals and the corresponding performance obtained by their multi-microphone version applied to all the far microphone signals.*

case, MPF performs better than the two other algorithms. A similar trend was obtained for position P2.

Moreover, as observed in our former activity on the development of multi-microphone WAUTOC, applying to far microphone signals any of the algorithms in single-channel fashion always led to a performance worse than that obtained using the MPF based algorithm.

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 (typical of an office environment), it is shown that the use of the proposed MPF 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 the single microphone version of the described algorithms to the output of each far acoustic sensor would in any case require a further processing to select the most reliable F0 among the resulting candidates.

Next steps include the application of MPF algorithm to the analysis of larger databases of meetings and lectures, collected and annotated under the CHIL project. The objective is to exploit the resulting F0 estimates as well as the MPF function as features for acoustic event detection and classification, speech activity detection, and eventually distant-talking ASR.

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] L. Armani, M. Omologo, "Weighted Autocorrelation-based F0 Estimation for Distant Talking Interaction with a Distributed Microphone Network", *Proc. of ICASSP*, 2004.

[4] 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.

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

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

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

[8] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components", J. Acoust. Soc. Am., Dec., 2004.

[9] ftp://ftp.cs.keele.ac.uk/pub/pitch, "Keele Database", University of Keele (UK).

[10] P. Boersma, D. Weenink, "Praat: doing phonetics by computer (Version 4.2, 2004)" [Computer program]. Retrieved from http://www.praat.org/