On the Use of a Weighted Autocorrelation Based Fundamental Frequency Estimation for a Multidimensional Speech Input

Federico Flego, Maurizio Omologo, Luca Armani

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

Abstract

The problem of computing the fundamental frequency F0 in an accurate way is a known and still partially unsolved problem, especially when dealing with a noisy speech input. In this work, a distant-talking scenario is addressed, where a distributed microphone network provides multi-channel input sequences to process for speaker modeling purposes. Given this context, one may process in an independent way each channel and then apply a majority vote or other fusion methods. Otherwise, the redundancy across the channels can be exploited jointly by processing the different signals to obtain a more reliable and robust F0 estimation. The paper investigates the use of a multi-channel version of a Weighted Autocorrelation (W AUTOC)-based F0 estimation technique. A postprocessing corrective step is introduced to improve the resulting F0 accuracy. Experiments conducted on a real database show the advantages and the robustness of the proposed method in extracting the fundamental frequency with no regard about the configuration of the microphone and talker position as well as the head orientation.

1. Introduction

An attractive future scenario consists of 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.

Towards this direction of ubiquitous computing, removing any constraint in the distribution of the microphones in space represents an important potential in terms of flexibility of the application, namely speaker tracking or distant-talking Automatic Speech Recognition (ASR).

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 him/herself. Although noticeable advances have been made during the last decade in speaker tracking and distant-talking ASR [1, 2], the existing prototypes are often based on the use of a microphone array, located in a predetermined position and characterized by a specific geometry. Moreover, even if the user is not encumbered anymore by hand-held or head-mounted microphones in most cases she/he can talk up to a limited distance from the microphones (a few meters), depending on the complexity of the environmental acoustics, and with constraints in the head orientation and in the speaking style.

In the PEACH project, a multi-channel scenario with an arbitrary microphone distribution is addressed as alternative to a traditional microphone array. Rather than focusing primarily on improving the quality of a spatial filtering process for enhancement purposes, we think to analyze the given acoustic scenario through a multi-channel processing aimed at extracting basic information to track talkers, classify acoustic events, and eventually recognize what has been uttered.

One way to pursue all these objectives is that of deriving a model of the source (e.g. the speaker) from the given multi-channel data. In the past, other attempts were made to extract a common source model from a microphone array processing. Brandstein et al. [1] contributed with a relevant work introducing the explicit-speech modeling. The basic concept is that each microphone signal represents a different observation of the same speech generation process, so that a common source model can be used in the related processing of all the channels. A common source model can be represented as a combination of common excitation, vocal tract and radiation models. Beside the source model, one may try to introduce characteristics of background noise and environmental acoustics (e.g. room impulse response) in this modeling framework, but this issue goes beyond the scope of this work.

In order to focus on the problem of deriving a common excitation model from the various input signals, as a first step, we are interested in computing the fundamental frequency F0 and, as a second step, a Voiced/UnVoiced or a more detailed excitation source classification, as ground for tracking natural variations in the characteristics of the vocal-tract system [3]. The simplest way to perform F0 estimation is to extend to the multi-channel case a paradigm that works for a single channel close-talking case. Hence, the Weighted Autocorrelation (W AUTOC) is here adopted [4]: it consists in an autocorrelation function weighted by the reciprocal of the Average Magnitude Difference Function (AMDF) [5], which turns out to be robust in processing noisy signals.

In [6], preliminary experiments showed a simple but effective method to extend the W AUTOC function to the multi-channel scenario. In this paper, a corrective procedure is adopted to further improve fundamental frequency estimation accuracy. To show the usefulness of the resulting technique, a small database [7] of speech sequences was reproduced in an office environment and recorded by a set of ten microphones distributed in space.

The paper is organized as follows: Section 2 introduces the distributed microphone network being investigated; Section 3 and Section 4 present the multi-channel W AUTOC method and the postprocessing corrective procedure, respectively; Section 5 describes the given experimental set-up; Section 6 refers to the evaluation criteria and Section 7 reports on the experimental results that were obtained.

This work was partially funded by the Province of Trento, under Fondo Unico PEACH project.
2. Distributed Microphone Network

The term “Distributed Microphone Network” is here introduced to describe a generic set of microphones localized in space without any specific geometry and connected to a recording and computing system, that ensures a sample-level synchronous processing of the corresponding signals.

As described in many references in the literature [1, 8], uniform linear arrays, harmonic arrays and large arrays are the most commonly investigated sensor configurations in multi-channel speech and audio processing. By applying related beamforming or spatial-temporal filtering techniques one can selectively pick-up a speech message and enhance it by reducing some of the undesirable effects due to distance, background noise, room reverberation and competitive sound sources. However, the geometry of the array may influence consistently its performance. In particular, the inter-microphone distance represents a key element in order to avoid the so-called spatial aliasing effect. If this distance does not fulfill specific requirements grating lobes are introduced at higher frequencies, that is, the array will pick-up interfering signals or reverberation components from directions other than the desired one.

On the other hand, in the scenario addressed by this work, the microphone network consists of several microphones or microphone pairs distributed in space at rather large distance (50 cm or more), which represents a multi-channel distribution that would introduce spatial aliasing at frequencies of interest for speech analysis and recognition, if combined with traditional beamforming techniques. An experiment described in [6] addressed the detrimental use of the resulting delay-and-add “beamformed” signal in the given experimental context. A secondary aspect also addressed in that paper regarded the time alignment among different microphone signals. In the following of this paper all the signals are assumed to be time aligned with each other. This was achieved by manually checking in order to avoid any possible offset delay error.

3. WAUTOC-based F0 estimation

In the past, many F0 (or pitch) estimation methods were proposed and evaluated [9, 10]. Some of these methods derive from the basic formulations of short-term autocorrelation and AMDF functions.

3.1. AMDF-based algorithms

The AMDF function is defined for a given time lag $\tau$ as follows:

$$\text{amdf}(\tau) = \frac{1}{N} \sum_{n=0}^{N-\tau-1} |x(n) - x(n + \tau)|,$$  \hspace{1cm} (1)

where $x(n)$ denotes the input signal and $N$ is the analysis frame length.

In case of a quasi-periodic input sequence, this function becomes small for a lag $\tau$ equal to the pitch period and, often, to its multiples. So, the common procedure consists in applying a rectangular window to block the signal into overlapping frames of length $N$ (e.g. $N = 800$ at 20KHz), and then calculate the AMDF for each frame and possible lag: for a given frame, the resulting period, in samples, is given by

$$l = \arg \min_{\tau} \{\text{amdf}(\tau)\}. \hspace{1cm} (2)$$

Moreover, the range where to find $l$ is limited by the fact that adult humans have a fundamental frequency between 50 and 400Hz.

3.2. Weighted Autocorrelation

Using the short-term autocorrelation instead of the AMDF function leads to similar F0 estimation methods.

The weighted autocorrelation-based method recently introduced by [4] proved to be particularly robust to noise and to doubling/halving period estimation mismatch. The WAUTOC function is defined as follows:

$$\text{waucoc}(\tau) = \frac{\sum_{n=0}^{N-\tau-1} x(n) x(n + \tau)}{\sum_{n=0}^{N-\tau-1} |x(n) - x(n + \tau)|} + \epsilon,$$  \hspace{1cm} (3)

where $\epsilon$ is a constant value that prevents the function from getting too high dynamics or zero-division condition.

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 estimation of the pitch period is derived as

$$l = \arg \max_{\tau} \{\text{waucoc}(\tau)\}. \hspace{1cm} (4)$$

3.3. Multi-channel WAUTOC

To extend the WAUTOC-based technique to the multi-channel case, a suitable way is that of averaging the given function over the entire microphone network, which leads to the computation of

$$f(\tau) = \sum_{i=1}^{M} w_i \cdot \text{waucoc}_i(\tau),$$  \hspace{1cm} (5)

where $M$ denotes the number of microphone channels, and the coefficients $w_i$ have been introduced to represent the reliability of each channel $\text{waucoc}_i$ function, which may depend on the speaker position and head orientation. As at this moment quantifying a channel reliability feature would not be trivial, in the following a constant value will be assigned to all of the coefficients ($w_i = 1/M$).

4. F0 correction

Several possible smoothing techniques may be applied to the F0 contour resulting from the application to a single channel input of one of the methods previously described. However, in the given multichannel framework, a simple smoothing approach can be firstly conceived taking into account both the F0 estimated from each channel in the current frame, and the pitch period value evaluated for the previous frame.

4.1. Median “filtering”

This approach simply considers the pitch period estimates for each channel $i$, computed for a given frame $j$,

$$l_{ij} = \arg \max_{\tau} \{\text{waucoc}_i(\tau)\}, \hspace{1cm} (6)$$

and returns a final pitch estimate, $l_j$, computed as follows:

$$l_j = \text{median}(l_{1j}, l_{2j}, \ldots, l_{Mj}), \hspace{1cm} (7)$$

where median(·) provides the middle value of the ordered list, which should be less sensitive to outlier values.
4.2. FFT-based correction

The second approach works in two steps. First a pitch correction is applied to each microphone channel, then the final pitch estimate provided by (5) is corrected by means of (7), in case a consistent change of pitch occurs from the previous frame.

\[
\frac{|l_{ij} - l_{ij-1}|}{l_{ij-1}} > \Theta_i
\]  

(8) 

the FFT is applied to the signal frame \( j \)-th to obtain the spectrum \( S_j(f) \).

Then, given the frequency pitch values \( F0_{ij} \) and \( F0_{ij-1} \), corresponding to \( l_{ij} \) and \( l_{ij-1} \), respectively, the first \( K-1 \) harmonics are considered for each of them. Afterwards, the sum of all the power FFT values associated to the generic \( F0 \) and its related harmonics is computed: the two resulting power sums are here denoted as \( \alpha_{ij} \) and \( \alpha_{ij-1} \).

In case \( \alpha_{ij} < \alpha_{ij-1} \), a new pitch period estimate is provided for the given channel by

\[
l'_{ij} = \arg \max_{\tau} \left\{ \text{wautoc}_{ij}(\tau') \right\},
\]  

(9) 

limiting \( \tau' \) in the range \( l_{ij-1} \pm \delta \).

According to our experience, this method guarantees that the pitch period is selected around the value provided by the previous frame, in case it seems to be more reliable than the current frame hypothesis.

On the basis of some preliminar analysis on a small evaluation set, \( \Theta, \delta \) and \( K \) were assigned the values of 0.2, 20 (i.e. 1ms) and 3.

Step 2: considering the final pitch estimates \( l_j \) and \( l_{j-1} \), for the current and previous frame, provided by

\[
l_j = \arg \max_{\tau} \left\{ \sum_{i=1}^{M} \text{wautoc}_{ij}(\tau) \right\},
\]  

(10) 

equation (8) is evaluated. In case a consistent variation is found, equation (7) is applied to the set of pitch period values obtained from Step 1, and the result, \( l'_j \), used as the new \( F0 \) estimate.

5. Experimental set-up

In order to measure the performance of the proposed algorithm, the Keele database was used [7], 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-channel database was needed, the Keele database was reproduced by using a very high quality dual-concentric (TANNØY 600A) loudspeaker, placed in two positions (P1 and P2) in order to have different sound propagation situations. Speech sequences were then recorded using 10 omni-directional microphones and a multi-channel platform operating at 20 kHz and 16 bit.

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 1, 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.

![Figure 1: 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 *](image)

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 [10]. 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.

The Keele database used in this work is available with laringograph data. \( F0 \) candidates are evaluated only for analysis frames labeled as “voiced” by the database human labeler.

A frequently used method to compare performances between different algorithms is to compute the Gross Error Rate (GRE). 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 percentages figures of 20% and 5% have been used for evaluation. In the literature, other common evaluation methods are based on reporting standard deviation of relative \( F0 \) errors or other types of relative deviation performance.

Another index used here for error estimation is the relative error rate which, for a given sequence of \( N_{fr} \) frames, is defined as

\[
\text{RelErr} = \frac{1}{N_{fr}} \sum_{n=1}^{N_{fr}} \frac{|F0(n) - F0_{w}(n)|}{F0_{w}(n)},
\]  

(11) 

7. Experimental results

In the following experiments, \( F0 \) estimates were obtained using an analysis step of 1 ms and an analysis window length of 30 ms. Hence, for voiced segments a rate of 1000 estimates/s was used for comparison purposes.

As shown in Table 1, a first experiment was conducted to derive the \( F0 \) estimation accuracy by directly processing the original database of close-talking signals. The Table shows that the WAUTOC technique leads to a relative error of 4.4%, which corresponds to gross error rates of 4.4% and 9.0%, and an average relative deviation of 3.6%.

The first experiment involving the use of far microphone channels refers to the application of WAUTOC-based method introduced in Section 3.3, that is with all the microphone signals processed with the same weights \( w_i \). Table 1 reports on the obvious increase that was observed in the relative error rates, from 4.4% to 8.1% and 11.1%, for positions P1 and P2, respectively.
Table 1: Error rates obtained by applying the WAUTOC-based method to the close-talking signal, and in its extension to the multi-channel case. +correction refers to results obtained applying methods described in Section 4.1 and 4.2, respectively.

<table>
<thead>
<tr>
<th></th>
<th>rel err [%]</th>
<th>GRE20 [%]</th>
<th>GRE5 [%]</th>
<th>avg rel dev [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>close talk</td>
<td>4.4</td>
<td>4.4</td>
<td>9.0</td>
<td>3.6</td>
</tr>
<tr>
<td>+correction</td>
<td>3.3</td>
<td>2.3</td>
<td>7.2</td>
<td>2.5</td>
</tr>
<tr>
<td>multichannel,P1</td>
<td>8.1</td>
<td>7.4</td>
<td>17.7</td>
<td>5.9</td>
</tr>
<tr>
<td>median</td>
<td>7.1</td>
<td>7.9</td>
<td>18.9</td>
<td>5.5</td>
</tr>
<tr>
<td>+correction</td>
<td>6.7</td>
<td>6.6</td>
<td>17.4</td>
<td>4.9</td>
</tr>
<tr>
<td>multichannel,P2</td>
<td>11.1</td>
<td>9.0</td>
<td>21.0</td>
<td>7.8</td>
</tr>
<tr>
<td>median</td>
<td>10.1</td>
<td>10.5</td>
<td>23.2</td>
<td>7.5</td>
</tr>
<tr>
<td>+correction</td>
<td>8.8</td>
<td>8.4</td>
<td>21.2</td>
<td>6.3</td>
</tr>
</tbody>
</table>

Figure 2: The two curves show the relative error rates derived by WAUTOC-processing each of the 10 microphones separately. Results refer to both loudspeaker positions. The three horizontal lines indicate the baseline for close-talking signals and the results obtained by applying the WAUTOC-based technique (position1-all, position2-all) to all the microphone signals.

However, as shown by Figure 2, if every far-microphone signal is processed by WAUTOC in the traditional single channel fashion, for all of the cases in position P1 (and almost all in P2) the relative error rate is worse than applying the multi-channel WAUTOC-based algorithm. The same trend of results can be observed for both loudspeaker positions. It is also worth noting that applying the multi-channel WAUTOC-based algorithm blindly is straightforward; on the other hand, developing a method to select the best microphone would be much more difficult.

For what concerns the application of the median filtering introduced in Section 4.1, one can note that an improvement is obtained in terms of relative error rate, if compared to the application of multichannel WAUTOC described in Section 3.3. However the technique does not provide the same improvement when the comparison is conducted in terms of Gross Error Rate. This fact is related to a bias in the distribution of pitch errors. However, when the median filtering is used together with the corrective procedure introduced in Section 4.2, results show an improvement, with respect to the baseline, for most of given evaluation criteria, as shown in Table 1.

It is worth noting that the corrective procedure provides an improvement also when F0 is estimated from the close-talk signal. This is due to the fact that a short history (one frame) is here exploited fruitfully.

8. Discussion and conclusions

This paper addressed the problem of estimating the fundamental frequency in real office environment, given a set of microphones distributed in space.

Although signals are degraded by noise and reverberation, it is shown that the use of a weighted autocorrelation method extended to a multi-channel synchronous processing allows to obtain better performance than using any of the microphone signal as single channel. In addition simple correction techniques have been investigated, that provided an improvement with respect to the baseline system.

Future work will be devoted to test the proposed methods on larger databases which will provide a better support for statistically based parameter estimation.

Another topic to address regards the development of a technique to extract other information about the excitation model, in particular voiced/unvoiced decision and pitch epochs in order to perform a pitch synchronous LPC or spectral analysis.

9. References