Evaluation of the pitch estimation algorithms in the monopitch and multipich cases

François Signol, Claude Barras, Jean-Sylvain Liénard

Evaluation of the Pitch Estimation Algorithms in the monopitch and multipitch cases

F. Signol, C. Barras and J.-S. Lienard

LIMSI-CNRS, BP133, 91403 Orsay Cedex, France
jean-sylvain.lienard@limsi.fr

Reliably tracking the fundamental frequency F_0 of the components is an important step in the separation of superimposed speech signals. Several Pitch Estimation Algorithms (PEAs) are potentially usable and a rigorous evaluation method is needed. However, even in the monopitch case, many variations between them render such a comparison difficult. The F_0min-F_0max interval extent, the use of a priori information on the whole sequence or database and above all the arbitrary voicing threshold setting lead to large differences in the results. These biases can be removed by setting the F_0 bounds to fixed values acceptable for many voices, by proceeding with the evaluation on a strictly frame-to-frame basis, and by fixing the voicing threshold in order to get an equal error rate for overvoiced and undervoiced frames. In the multipitch case any frame may exhibit 0, 1 or 2 valid voicing according to the coincidence between the voiced and unvoiced parts of both signals. This problem is treated by defining a metric linking the PEAS's hypotheses to the pitch values of the isolated signals. The proposed methodology is applied to several PEAs on several databases in the monopitch case.

1 Introduction

In the last five decades, numerous PEAs have been developed so that a proper evaluation is mandatory. Most PEAs process the input speech signal as successive short-time frames. A F_0 estimate is produced for each voiced frame. For unvoiced frames the F_0 estimate is replaced by a zero value. This involves two distinct processes: Voiced/Unvoiced (VuV) frame decision and F_0 estimation. In some cases those processes are performed sequentially. In [1,3] the first step is the computation of a voicing criterion (typically between 0 and 1). Then this voicing criterion is compared to a threshold and the Voiced/Unvoiced (VuV) decision is taken. This is a typical sequential frame-to-frame approach. In some other cases the two processes are not distinct and sometimes they are mutually dependent. For instance in [2,6] several F_0 candidates are selected for each frame and a dynamic programming algorithm is used to select the globally optimal sequence of candidates. In this case, the F_0 estimation influences the VuV decision and vice-versa. Eventually evaluating a PEA means simultaneously evaluating the two above processes: VuV decision and F_0 estimation.

PEAs evaluation requires a database, VuV references, F_0 references and quality measures. Thus several variability sources should be addressed, which means numerous features to set and control. First of all, one has to control the database specificities such as the speech recording conditions (noise level, anechoic room, telephone, car...) or the type of speech (read, broadcast news, conference, spontaneous, expressive...). The reference voicing and F_0 labeling has to be perfectly reliable. It may differ according to the protocol used (manual, fully automatic or manually corrected); it may come from the analysis of the speech signal itself, or from an electroglottographic waveform (EGG) recorded simultaneously. The VuV decision may have been taken on the basis of the experience of trained phoneticians, or from the examination of a waveform on physical criteria, which may be quite different. Finally each PEA's specificity should be taken into account. It may or may not offer the operator the capability of disabling some pre- or post-processing, or of modifying the main parameters (frame duration, frame hop, voicing threshold, lower and upper bounds of F0 estimation, number of F0 candidates per frame). For all the above reasons, it is extremely difficult to perform a fair comparison between algorithms.

In this paper we propose an evaluation methodology that could help containing some of the main sources of variability. Section 2 presents what has been done in PEA's evaluation, highlights how the VuV decision is problematic to evaluate and proposes a way to deal with it. The usual PEA use is the monopitch case where it has to deal with a one-speaker speech signal. Section 3 proposes a new evaluation methodology and illustrates it on a database in the monopitch case. Section 4 extends our monopitch evaluation approach to the multipitch case, in which several voices may be mixed in a single signal.

2 The role of the VuV decision in the evaluation of F_0 estimation

The first quality measure for which a consensus in literature exists is the quality of the pitch estimation. What is generally reported is a gross error rate (GER). A gross error is raised when the pitch estimate lies more than a certain distance away from the reference (typically 20%). This rather high threshold has to be put in perspective with the typical pitch errors which are mainly octave or sub-octave errors and most seldom third, triple or two thirds errors [5]. The gross error rate does indeed capture the principal harmonic or sub-harmonic errors.

F_0 estimation is performed on the frames which are considered as voiced. In that sense, the GER calculation is subjected to the VuV decision. However, the voicing state may be ill-defined for many frames, often located at the bounds of the voiced speech segments. In those regions F_0 estimation may remain unsteady, whereas for frames where the voicing is strong, the F_0 estimation is quite reliable. For any PEA, avoiding the litigious frames may drastically reduce the F_0 error rate. This phenomenon is illustrated on Fig. 1, which shows that the GER decreases and falls to 0% when the VuV threshold is increased so that only the reliably voiced frames are taken into account. This example was performed on a short sample of speech with the YIN algorithm, but the same trend can be observed with any algorithm, on any speech excerpt.

Therefore, for a given PEA the GER calculation should not be presented alone. It should be linked with an evaluation of the VuV decision.

The results published in [3] are established on the frames declared as voiced by all PEAs involved in the evaluation. Thus they discard from the statistics the frames whose voicing is not reliably established. This protocol may reflect the intrinsic quality of the algorithm, in dealing with perfectly voiced sounds. But it does not give a fair idea of the PEA's behavior in the difficult situation encountered at both ends of the voiced segments. Thus, we have to normalize the algorithms behaviors in terms of VuV decision.

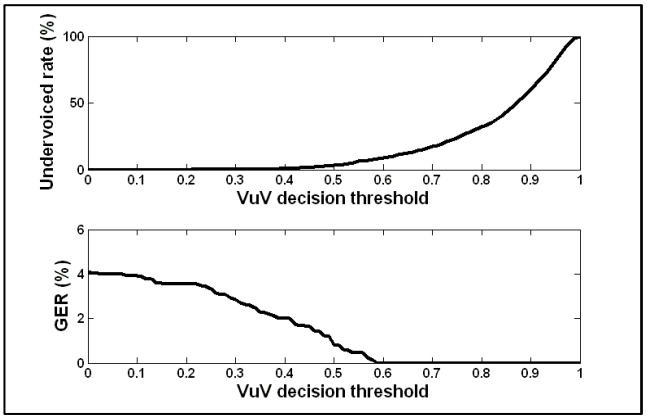

Fig. 1. Rate of frames declared as voiced and Gross Error Rate as functions of the VuV decision threshold

To formalize our voicing behavior normalization let us look at table 1 which illustrates the kinds of errors that could appear between the VuV state of a reference frame and the VuV state of an hypothesis frame. A false alarm (FA) called overvoiced error occurs when the hypothesis frame is voiced and not the reference frame. A false rejection (FR) called undervoiced error occurs in the opposite situation.

Reference	Hypothesis	Agreement
UV	UV	OK
UV	V	FA
V	UV	FR
V	V	OK

Table 1. Full set of reference/hypothesis couples and their agreement result.

Table 1 refers to a typical problem of decision theory. This problem may be addressed within the Equal Error Rate (EER) statistical framework. A way to compare the different PEAs without the bias attached to their specific VuV behavior is to tune it so that the FA/FR ratio gets a constant value. Let EER be the ratio of the overvoiced error rate to the undervoiced error rate. The value chosen for the EER depends on the voicing behavior requested by the application where we want to use a PEA. For instance, if the application is only interested in strongly voiced segments, we should set EER at a value between 0 and 1 in order to get more undervoiced than overvoiced frames. For the present evaluation purpose we shall fix it at the value 1.

3 Monopitch case

In this chapter we present our evaluation methodology beginning with the database presentation. Then we explain precisely how we build our reference. Then we provide a general formalization of how could be represented a set of reference annotations and F_0 hypotheses. At last, we propose some evaluating features and formalize them in the monopitch case.

3.1 Evaluation method

3.1.1 Database

Our mono-speaker database is composed of three speech corpora, all recorded in clean acoustic conditions. The Keele database (noted K) comprises 10 English speakers (5 males, 5 females) reading the same text one time and quite neutrally, for a total duration of 337 seconds. The Bagshaw database (noted B) comprises 2 English speakers (1 male, 1 female). Both are saying the same 50 shorts utterances, for a total of 331 seconds. The Daless database (noted D) comprises 4 French speakers (2 males, 2 females) reading various texts quite neutrally. This database contains 1359 seconds of speech. For all the databases, the authors give the EGG waveform for each speech signal.

To build a 2-speaker mixture, two normalized signals are simply added. Then, the whole 2-speaker mixtures databases are built by forming all possible signals combinations.

3.1.2 Reference annotations

This step needs a particular care because it determines the quality of the evaluation results. It raises the “groundtruth” problem.

Manual or manually corrected reference annotation is a very tedious task. It must be performed by a group of specialized operators to be reliable. However, the protocols and groups of operators differ from one database to the next, so that it may happen that a given algorithm gets different results when evaluated on different databases. This is why we chose to automatize the annotation process, even if we know that it may produce more errors than the manual annotation of a given database. Also, as there are few manually annotated databases, this approach permits the use of a non-annotated database. For instance our largest database is not manually annotated; an automatic annotation procedure was the only way to use it with the same annotation quality than the other two. In any case, a careful examination of each database is needed in order to adapt some parameters such as the F_0 range.

As EGG signals are available for all databases, we generate the VuV decision with them. Although it is not perfect, the EGG waveform is a direct trace of the vocal cords vibration so it should give a more reliable and steady voicing measure than the acoustic speech signal.

The EGG waveform is first shaped by a bandpass filtering between 50 and 1600 Hz to cancel undesired noises and low-frequency components. Then, a positive saturation threshold is computed which corresponds to the amplitude exceeded by 5% of the samples. The same negative saturation threshold is computed with the negative part of the waveform. If the negative saturation threshold's absolute value is bigger than the positive saturation threshold then only the absolute value of the waveform's negative part is kept (negative half-wave rectification). Else a positive half-wave rectification is done. The obtained waveform is normalized at 0.9 and then the temporal envelope is computed by linear interpolation between local maxima. A 18-point histogram is computed from the envelope. The histogram exhibits two peaks, corresponding respectively to the unvoiced and to the voiced segments. The local minimum between these two peaks is chosen as the voicing threshold. Below this threshold the sample is considered unvoiced and above the sample is voiced. A refinement is added to avoid too short voiced or unvoiced segments. All unvoiced segments shorter than a first

threshold and surrounded by voiced segments become voiced segments (voiced fusion). Then the isolated voiced segments shorter than a second threshold become unvoiced.

This algorithm returns a series of voiced segments defined by a beginning time and a finishing time. A frame is voiced if at least half of its width is included in a voiced segment.

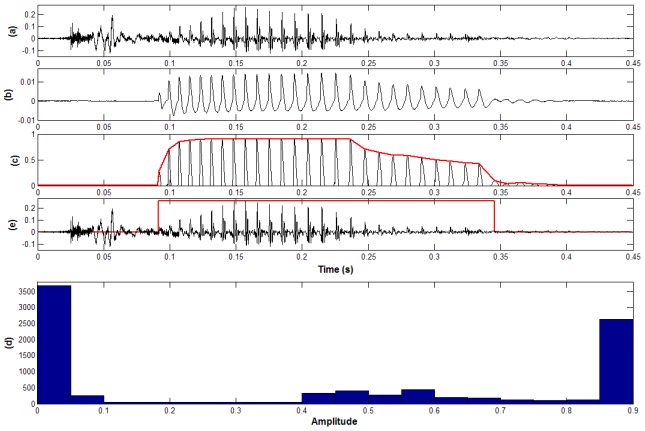

Fig 1. (a) speech signal (b) egg signal (c) filtered, saturated, half-wave rectified, normalized egg (black line) and temporal envelope (red line) (d) speech signal (black line) and voiced segment (red line) (e) envelope 18-points histogram. The voicing threshold is 0.1.

To compute the F0 reference value for a voiced frame, a basic and easily reproducible algorithm is used. The EGG signal is windowed. On each window, normalized autocorrelation is computed, its first main lobe is removed and the abscissa of the highest peak is considered as the period of the fundamental for this frame.

This algorithm returns a series of frames described by two values: the frame's time which is the middle instant of the EGG window studied, and the estimated F_0 which is strictly positive if the frame is voiced and 0 else.

3.1.3 Quality measures

Let us define some notations.

We note N_S the number of speakers mixed in the speech signal. N_R corresponds to the number of F_0 values in each reference frames. N_R is equal to N_S because there are as much references values as speakers. N_H is the number of F_0 hypothesis candidates given by a PEA per frame. N_H is superior or equal to N_R. We also note N_F the total frame's number.

Let us note $\cap$ the intersection operator, $\cup$ the union operator, Ω the cardinal operator and | the writing shortcut which signifies "such as".

We define R as the references frames set and R^t as the t^{th} reference frame. R^t contains the set of F_0 references values denoted R_x^t. R_x^t is the x^{th} F_0 reference value in the t^{th} reference frame. The same kind of notation is applied to the hypotheses. H is the set of hypotheses frames. H^t is the t^{th} hypotheses frame and H_y^t is the y^{th} F_0 hypothesis value in the t^{th} hypotheses frame. The range of t, x and y are given in Eq.(1).

$$\begin{matrix} R=\{R^1 \dots R^{N_F}\} \\ H=\{H^1 \dots H^{N_H}\} \end{matrix} \text{ and } \begin{matrix} R^t=\{R_1^t \dots R_{N_R}^t\} \\ H^t=\{H_1^t \dots H_{N_H}^t\} \end{matrix} \quad (1)$$

An unvoiced F_0 candidate is set to zero. Else, it is strictly positive. We note VC_R the set of voiced F_0 references values and uVC_R the set of unvoiced F_0 references values. VC_H is the set of voiced F_0 hypotheses values and uVC_H the set of unvoiced F_0 hypotheses values. A mathematical formulation is given in Eq.(2).

$$\begin{matrix} VC_R=\{(x,t)|R_x^t>0\} \\ VC_H=\{(y,t)|H_y^t>0\} \end{matrix} \text{ and } \begin{matrix} uVC_R=\{(x,t)|R_x^t=0\} \\ uVC_H=\{(y,t)|H_y^t=0\} \end{matrix} \quad (2)$$

Concerning the VuV decision, we consider a frame as voiced if at least one of its F_0 values is strictly positive. Thus let us note (R^t>0) as a voiced references frame and (H^t>0) as a voiced hypotheses frame. We also note (R^t=0) an unvoiced references frame and (H^t=0) and unvoiced hypotheses frame. These definitions are illustrated in the Eq.(3).

$$\begin{matrix} R^t>0 \Leftrightarrow \exists x|R_x^t>0 \\ H^t>0 \Leftrightarrow \exists y|H_y^t>0 \end{matrix} \text{ and } \begin{matrix} R^t=0 \Leftrightarrow \forall x|R_x^t=0 \\ H^t=0 \Leftrightarrow \forall y|H_y^t=0 \end{matrix} \quad (3)$$

VF_R is defined as the set of voiced references frames and VF_H is the set of voiced hypotheses frames. uVF_R is the set of unvoiced references frames and uVF_H the set of unvoiced hypotheses frames.

$$\begin{matrix} VF_R=\{t|R^t>0\} \\ VF_H=\{t|H^t>0\} \end{matrix} \text{ and } \begin{matrix} uVF_R=\{t|R^t=0\} \\ uVF_H=\{t|H^t=0\} \end{matrix} \quad (4)$$

We define two series of quality measures: the first one deals with the VuV decision, and the second one deals with F_0estimation.

In the monopitch case, N_R and N_S are equaled to 1 and the number of references frames is the same than the number of F_0 references values.

The overvoiced frame rate (OVR) corresponds to the number of FA frames over the number of unvoiced references frames as explained in Eq.(5). The undervoiced frame rate (UVR) is the ratio between the FR frames number and the number of voiced references frames. It is given in Eq.(6).

$$OVR=100\frac{\Omega[uVF_R \cap VF_H]}{\Omega[uVF_R]} \quad (5)$$

$$UVR=100\frac{\Omega[VF_R \cap uVF_H]}{\Omega[VF_R]} \quad (6)$$

The voiced decision agreement rate (V_{ok}) is given in Eq.(7). It corresponds to the number of common voiced frames between the reference and the hypothesis over the number of voiced frames in the reference or the hypothesis. The unvoiced decision agreement rate (uV_{ok}) is described in Eq.(8).

$$V_{ok}=100\frac{\Omega[VF_R \cap VF_H]}{\Omega[VF_R \cup VF_H]} \quad (7)$$

$$uV_{ok}=100\frac{\Omega[VF_R \cap VF_H]}{\Omega[VF_R \cup VF_H]} \quad (8)$$

The F_0 estimation quality measure involves a Gross Error Threshold (GET) fixed at 20%. A F_0 reference value is said "in accordance" with a F_0 hypothesis value whether their absolute relative distances is inferior to GET. Eq.(9) presents the mathematical formalization of this definition. We note the "in accordance" operator by the ≈ symbol.

$$R_x^t \approx H_y^t \Leftrightarrow \left(R_x^t > 0\right) \wedge \left(\exists y \left| \frac{\left| R_x^t - H_y^t \right|}{R_x^t} \leq GET \right. \right) \quad (9)$$

The "in accordance" operator is extendable to frames. A reference frame is in accordance with a hypothesis frame if all the F_0 references values are in accordance as showed in Eq.(10).

$$R^t \approx H^t \Leftrightarrow \forall x, \exists y \left| R_x^t \approx H_y^t \right. \quad (10)$$

The set of references frames in accordance is noted R_{ok} and is described in Eq.(11).

$$R_{ok} = \left\{ t \left| R^t \approx H^t \right. \right\} \quad (11)$$

The "in accordance" rate corresponding to a recall rate is noted RER and is given by Eq.(12). It is the ratio between the number of "in accordance" reference frames and the number of common voiced frames between reference and hypothesis.

$$RER = \frac{\Omega\left[R_{ok}\right]}{\Omega\left[VF_R \cap VF_H\right]} \quad (12)$$

The precision rate (PRR) corresponds to the number of "in accordance" reference frames over the number of voiced F_0 hypotheses values needed to put in accordance the reference frame. It is an indication of the PEA efficiency to find the rights F_0 values in the first hypotheses candidates. PRR is explained in the Eq.(13) below.

$$PRR = \frac{\Omega\left[R_{ok}\right]}{\sum_{y \in [1, N_H], t \in H_{ok}} \Omega\left[\left\{H_y^t > 0\right\}\right]} \quad (13)$$

In the case of PEAs which provide a single F_0 hypothesis value per frame, PRR always equals to 100%.

The Gross Error Rate (GER) is theF_0 estimation accuracy indicator. It corresponds to the number of frames not "in accordance" over the total number of voiced references frames. We note R_{ko} the set of reference frames not "in accordance". The GER is given in Eq.(14).

$$GER = \frac{\Omega\left[R_{ko}\right]}{\Omega\left[VF_R\right]} \quad (12)$$

An hypothesis F_0 can be associated with several reference F_0 values. Thus if some F_0 reference values are equal in a same frame (crossing streams of F_0), PEAs probably return a single hypothesis for the ensemble of equal F_0 reference values. This multi-association allows the evaluation to deal with this problem.

Once a hypothesis F_0 value is associated to one or more reference F_0 values, the hypothesis F_0 value is locked and can not be associated with other reference F_0 values.

In the monopitch case, a F_0 candidate, a frame and a stream are the same. Thus, it remains only four voicing quality measures.

3.2 Monopitch evaluation results

Four PEAs are evaluated and compared: YIN, SWIPE, PRAAT and PSH. YIN [1] provides an aperiodicity criterion which can be easily converted into a voicing feature between 0 and 1. It gives one hypothesis per frame. SWIPE [3] provides a voicing strength criterion between 0 and 1 and also gives a single hypothesis per frame. PSH [4] provides a voicing strength which needs to be converted in a voicing feature between 0 and 1. It gives a tunable number of hypotheses and is adapted to the multipitch case. In the monopitch case, we fixed it at 1 like the others tested PEAs. PRAAT [2] provides an autocorrelation coefficient between 0 and 1 as voicing feature. In our series of algorithms, it is the only PEA with a post-processing. We tuned its parameters to remove this post-processing effect. All PEAs are tuned to provide a single F_0 hypothesis value so in Table 1 the PRR column is removed because PRR is always equal to 100%.

The particularities of each PEAs may lead to some small time misalignments between the references and the hypotheses so that a time alignment step is needed. It consists in a linear interpolation between the first before hypothesis value and the first after value. There is an undefined interpolation in the case of an unvoiced hypothesis frame and a voiced one. These frames are discarded from evaluation. As the reference F_0 range differs from the hypothesis F_0 range, all the F_0 references out of the hypothesis F_0 range are discarded from the evaluation.

	EER	V_{ok}	uV_{ok}	OVR	UVR	RER	GER
YIN	1.00	76.5	74.7	12.3	12.4	93.9	5.1
PSH	1.01	85.6	85.7	6.3	6.3	95.5	4.2
SWIPE	0.98	80.1	77.8	10.3	10.4	96.4	3.0
PRAAT	1.00	62.2	60.6	17.1	16.7	94.1	3.7

Table 2. Evaluation results on all the corpora.

For a given PEA, all statistics are means obtained on the whole of the three corpora. The results given in table 2 do not aim at a classification of the algorithms tested. They show that the performance differences are not as large as they look in the original publications, and they may encourage a detailed analysis of their respective strengths and weaknesses. The PRAAT settings that we adopted to remove the post-processing are clearly not adapted to this algorithm.

4 Multipitch case

In the multipitch case, several F0 streams need to be tracked simultaneously. This complicated task is usually performed by top-down approaches. These approaches use continuity hypotheses from one frame to the next or other general knowledge concerning the whole sequence. As there is also a lack of consensus in literature concerning these top-down processes, it is still difficult to evaluate their contributions to voicing and pitch estimation. That is why we are interested in a frame-to-frame evaluation without knowing more than the information extracted from the studied frame. We are aware of the fact that the better are the frame-level feature estimators, the simplest will the higher-level processes be.

4.1 Evaluation method

4.1.1 Database an reference annotations

To build a 2-speakers mixture, two 60 dB energy normalized signals are simply added. The energy normalization method is the "Scale Intensity..." PRAAT

function. Then, the 2-speakers mixtures databases are built by adding several possible signals combinations. The mixtures are cut to the shortest component. Three types of mixtures are available: the female/female mixtures, the male/male mixtures and the male/female mixtures. K contains 1411 seconds of speech mixtures, B contains 13693 seconds of speech mixtures, D contains 79869 seconds of speech mixtures for a total of 26 hours 22 minutes of speech mixtures artificially collected.

The reference annotations in the multipitch case are directly based on the monopitch reference annotations. The annotations consist in a simple concatenation of the monopitch reference annotations.

4.1.2 Quality measures

All the notations and definitions posed in 3.1.3 are still available in this section. In the monopitch case, a speech signal contains only one speaker so that it was not necessary to introduce the notion of F_0 stream. A F_0 stream corresponds to the set of F_0 values belonging to one speaker in a speech signal. In the multipitch case we have to introduce it due to the mixing of F_0 streams. One defines S_R the set of F_0 streams contained in the signal and S_R^S the particular F_0 stream of the s^{th} speaker. There are as many F_0 streams as speakers so there are N_S F_0 streams. An ideal PEAs should be able to reproduce the exact references annotations on his hypotheses output.

$$S_R = \left\{ S_R^1 ... S_R^{N_S} \right\} \text{ and } S_R^S = \left\{ R_S^1 ... R_S^{N_F} \right\} \qquad (13)$$

An F_0 hypothesis value should be associated with several reference F_0 values. Thus if some F0 reference values are equal in a same frame (crossing streams of F_0), PEAs probably return a single hypothesis for the ensemble of equals F_0 reference values. This multi-association allows the evaluation to deal with this problem. Symmetrically once a F_0 hypothesis value is associated to one or more F_0 reference values, the hypothesis F_0 value is locked and can not be associated anymore.

With the VuV definition adopted in 3.1.3, the VuV decision evaluation criteria remain the same in the multipitch case than in the monopitch case. One frame is voiced if one of her value is strictly positive. Mathematical definitions of V_{ok}, uV_{ok}, OVR an UVR remain unchanged. Nevertheless as the reference change, a new adequate VuV decision threshold has to be computed.

There is an evolution in the F_0 estimation evaluation criteria. Contrary to the monopitch case where a F_0 reference value is the same than a frame reference, in the multipitch case we can clearly distinguish between them and any simplification is no longer available.

The "in accordance" F_0 references values rate RECR given in Eq.(15) is the number of F_0 references values in accordance over the number of voiced F_0 reference values (VC_R defined in Eq.(2)). Eq.(14) provides the definition of RC_{ok} which is the set of F_0 reference values in accordance.

$$RC_{ok} = \left\{ (x,t) \middle| (\exists y \middle| R_x^t \approx H_y^t) \right\} \qquad (14)$$

$$RECR = \frac{\Omega[RC_{ok}]}{\Omega[VC_R \cap VC_H]} \qquad (15)$$

One can define the same rate for frames. We call REFR the "in accordance" reference frames rate. It corresponds to the number of reference frames in accordance over the number of voiced references frames. It is a first feature to quantify a PEA quality in reconstructing F_0 streams. Indeed higher is REFR higher is the number of frames where all F0 references values are associated with an F0 hypothesis value and better may be the F_0 streams tracking. Eq.(16) and (17) formalize this feature.

$$RF_{ok} = \left\{ t \middle| R_x^t \approx H_y^t) \right\} \qquad (16)$$

$$REFR = \frac{\Omega[RF_{ok}]}{\Omega[VF_R \cap VF_H]} \qquad (17)$$

The precision rate on F_0 values explained by Eq.(18) is the same than in Eq.(13) except that here it may exists more than one F_0 reference values.

$$PRR = \frac{\Omega[RC_{ok}]}{\sum_{t \in VF_R \cap VF_H} \Omega[\{H_y^t > 0\}]} \qquad (18)$$

5 Conclusions

In this paper we provided an evaluation methodology to compare PEAs not only on their F_0 estimation efficiency but also on the VuV decision which is a PEA step at least as important as the F_0 estimation. We illustrated this view in the monopitch case, by using several PEAs on a set of 3 different databases including the EGG signals, for which an automatic annotation protocol has been worked out. The methodology has been elaborated to extend easily to the multipitch case.

This work reinforces the idea that F_0 estimation and VuV decision are deeply related, in complex ways. Voicing was considered here as a two-state, binary problem. In the future it may become necessary to treat separately two notions of voicing, one at the signal level, essentially continuous, and the other, binary, incorporating some upper-level perceptive and linguistic considerations.

References

[1] de Cheveigné A., Kawahara, H., "YIN, a fundamental frequency estimator for speech and music", *J. Acoust. Soc. Am.* 111, 1917-1930 (2002).

[2] Boersma P., "Accurate Short-term Analysis of the Fundamental Frequency and the Harmonics-to-noise Ratio of a Sampled Sound", *IFA Proceedings* 17, 97-110 (1993).

[3] Camacho A., "SWIPE: a Sawtooth Waveform Inspired Pitch Estimator for Speech and Music", *Ph. Dissertation*, University Of Florida (2007).

[4] Liénard J-S., Signol F., Barras C., "Speech Fundamental Frequency Estimation Using the Alternate Comb", *INTERSPEECH 2007*, 2273-2276 (2007).

[5] Liénard J-S., Barras C., Signol F., "Using sets of combs to control pitch estimation errors", *ACOUSTICS 2008*, (2008).

[6] Secrest, B., Doddington, G., "An integrated pitch traking algorithm for speech systems", *Proc. ICASSP 1983*, 1352-1355 (1983).