# A PITCH DETERMINATION AND VOICED/UNVOICED DECISION ALGORITHM FOR NOISY SPEECH

Jean Rouat, Yong Chun Liu and Daniel Morissette
e-mail: jrouat@uqac.uquebec.ca

Université du Québec à Chicoutimi
ERMETIS, Département des sciences appliquées
555, boulevard de l'Université
Chicoutimi, Québec, Canada
G7H 2B1

**ABSTRACT**

We propose a multi-channel pitch determination algorithm (PDA) that has been tested on three speech databases (0dB SNR telephone speech, speech recorded in a car and clean speech) involving fifty-eight speakers. The system has been compared to AMPEX [9], to hand-labelled and laryngograph pitch contours. Our PDA comprises an automatic channel selection module and a pitch extraction module that relies on a pseudo-periodic histogram (combination of normalised scalar products for the less corrupted channels) in order to find pitch. It outperformed the reference system on 0dB telephone and car speech. The automatic selection of channels was effective on the very noisy telephone speech (0dB) but not on the car speech where the robustness of the system is mainly due to the pitch extraction module in comparison to AMPEX. The paper reports in details the V/UV, UV/V performance and pitch estimation errors for the PDA and the reference system on the three databases.

## 1. INTRODUCTION

The automatic tracking of pitch has many applications in the field of speech processing and speech technologies. One could enumerate many potential applications based on the automatic determination of pitch. But most of them are limited to clean speech and can not be used for real life applications due to the difficulty of determining pitch in adverse environment. The paper proposes a new PDA system and evaluates the performance in real and artificial noisy situations.

The automatic determination of pitch is one of the most difficult task in speech processing. Many pitch determination algorithms have been proposed and very few of them seem to work properly in a noisy environment. The pitch tracking of speech spoken in adverse environment is a challenging and yet unsolved issue that includes the difficulties from "standard" PDA (first formant close to the fundamental, speaker variability's, etc.) and combines those of obtaining the information out of the noise or of the interference. Among the recent pitch determination algorithms one can refer to the works by Van Immerseel and Martens [9] and Bagshaw et al. [1].

Van Immerseel and Martens [9] propose a pitch and voiced/unvoiced determination algorithm (called AMPEX), based on an auditory model. AMPEX is reported to be one of the best evaluated PDA, as the database is comprised of 14 males and 14 females speakers for a total speech duration of 56 seconds.

Bagshaw et al. [1] propose an enhanced super resolution pitch determinator (eSRPD) based on the work by Medan et al. [5]. The eSRPD was compared to six others systems : cepstrum pitch determination, feature-based pitch tracker, harmonic product spectrum, integrated pitch tracking algorithm, parallel processing method and super resolution pitch determinator. Bagshaw shows that eSRPD is the best one.

We compare the performance with the AMPEX system by using the same speech databases. The original AMPEX system has been kindly provided by Luc Van Immerseel and Jean-Pierre Martens from Gent University, Belgium, in order to compare both systems.

## 2. THE REFERENCE PDA (AMPEX)

### 2.1.The preprocessor

AMPEX (Auditory Model-based Pitch Extractor) [9] is a pitch extractor based on an auditory model. It comprises a middle ear filter, a bank of 20 auditory filters, a model of the mechanical to neural transduction and of auditory nerve transmission, plus a model of virtual tone component separation from the roughness and loudness. The PDA estimates pitch based on the virtual tone component by performing a pseudo-correlation analysis for each channel.

A global pseudo-autocorrelation function R(m) is obtained by accumulating the pseudo-autocorrelation across the channels.

### 2.2. The pitch extraction and the voiced/unvoiced decision

AMPEX searches for peaks in R(m), and for each peak that is larger than a threshold $\delta_r$, a pitch candidate is generated. The final pitch $T_0(n)$ and its evidence $E_0(n)$ for frame n are derived from pitch candidates generated for frames n-2 to n+2. If no candidate exists, then $T_0 = 0$ and its evidence is equal to zero. There is a continuity between two pitch candidates, $T_1$ and $T_2$, if

$$\left| \frac{T_1 - T_2}{T_1 + T_2} \right| \leq \delta_T \qquad (1)$$

where, $\delta_T$ is a threshold defined by the user. The voiced/unvoiced decision is based on the pitch evidence and on the continuity of the pitch estimates. A frame is unvoiced unless $E_0 > \delta_V$, or $E_0 > \delta_V/2$ and there is a continuity between $T_0(n)$ and $T_0(n-1)$.

According to Van Immerseel and Martens [9] there are two parameters that need to be optimised : $\delta_T$ and $\delta_V$; $\delta_r$ is not critical.

### 2.3 Modifications

The original AMPEX system has been adapted to our

needs (8kHz sampling rate, instead of 10kHz) by Luc Van Immerseel. When needed, we modified the original threshold parameters to improve the AMPEX performance.

## 3. THE PROPOSED PITCH DETERMINATION ALGORITHM

The auditory system is able to perceive a residue pitch even if the fundamental frequency is absent from the speech signal. The work by Delgutte [2] confirms that some auditory fibers show modulation patterns corresponding to harmonic interactions. Therefore, the auditory system is able to track pitch by relying on patterns of modulation for fibers influenced by a summation of stimulus harmonics [2] [6]. Such modulations are observed at the output of a perceptive filter-bank (see Patterson, [7] or Rouat et al., [8] for examples). The proposed model relies on the modulations at the output of a perceptive filter-bank to determine the pitch that is available through the channels.

The speech has been sampled to 8 kHz after proper low-pass filtering and down-sampling (depending on the database).

### 3.1. The auditory filterbank

The auditory filterbank simulates twenty groups of inner hair cells and covers the frequency range from 330 Hz to 4000 Hz. The dynamic compression properties of the inner hair cells and the exact nonlinearities of the mechanical to neural transduction are not taken into account in the present work as we do not intend to model exactly the peripheral auditory system.

### 3.2. The spectro-temporal module

#### *3.2.1 Low-frequency channels*

Channels one through eleven and channels twelve through twenty are subject to different processing. Channels one through eleven are low-pass-filtered. Then, the very low frequency distortions are removed by high-pass filtering to 63 Hz.

Thereafter, the normalised scalar products $R_i(\tau)$ between the vector $X_i(j)$, from channel i, and its time shifted version $Y_i(j)=X_i(j+\tau)$ is calculated for i = 1,...11.

$$R_i(\tau) = \frac{1}{\sqrt{E_X}} \frac{1}{\sqrt{E_Y}} \sum_{n=-N/2}^{N/2} x(j+n)x(j+n+\tau) \; ; \tau=13,...2N \qquad (2)$$

with $E_X = \sum_{n=-N/2}^{N/2} x(j+n)^2$ ; $E_Y = \sum_{n=-N/2}^{N/2} x(j+n+\tau)^2$ (3)

$E_X$ is the present segment energy, $E_Y$ is the energy of the shifted segment. The scalar product is calculated for pitch frequencies belonging to the interval [62.5Hz, 600Hz].

#### *3.2.2.Medium and high-frequency channels*

The output signals of channels 12 to 20 are processed by the Teager Energy operator [3][4]. Then, Teager Energy operator is filtered by performing exactly the same processing (low-pass and high-pass filtering) as described in section 3.2.1.

#### *3..2.3. Channels selection for the medium and high frequencies*

For a fixed speech frame and for i = 12,...20; do the following:

1. The energy $E_X$ of the current signal segment $X_i(j)$ for channel i, is compared with a reference energy, (energy from the previously non selected segment, for channel i).
2. If $E_X$ does not exceed the reference energy and if the previous speech segment has not been judged voiced by the Pitch Extraction Module (section 3.4),

Then

Do not select current channel, update the reference energy;

Else

Begin

Compute $R_i(\tau)$;

Find $T_{max}$= argmax ($R_i(\tau)$), (i.e. max($R_i(\tau)$) = $R_i(T_{max})$);

Compute a new normalised scalar product with a window length of 30 ms (instead of 16 ms for $R_i(\tau)$) and only for time lags, τ belonging to the interval [$T_{max}$-5, $T_{max}$+5].

If there is still a local maximum in this new product then, select the current channel, otherwise, do not select and update the reference energy.

End

3. Complete the process for all channels.
4. Stop.

#### *3.2.4. Channel combinations*

A new "global" scalar product is obtained by summing all the $R_i(\tau)$ over all the low-frequency channels and the selected medium high frequencies channels yielding:

$$PPH(\tau) = \frac{1}{M} \sum_{i=1}^{M} R_i(\tau) \qquad (4)$$

### 3.4. The pitch extraction analysis

The two largest peaks among the "valid" peaks from PPH(τ) are selected. To be "valid", a peak has to be greater than a fixed threshold S. If there is less than two "valid" peaks in the PPH, the segment is declared unvoiced. The algorithm searches for the sub-multiples of $\tau_1$ and $\tau_2$ and the pitch period T is the lowest common sub-multiple with PPH(T) being, at least, one of the first tenth highest "valid" peaks with

$$PPH(T) \geq S_{pe} . Max\left(PPH(\tau_1); PPH(\tau_2)\right) \qquad (5)$$

where $S_{pe}$ = 0.5. If the algorithm fails to find such T, the segment is declared as being "possible unvoiced", otherwise it is declared as being voiced and the pitch frequency is equal to 1/T.

### 3.5. Voiced, unvoiced decision

It is assumed that voiced segments found by the pitch extraction analysis are really voiced. Therefore, when a "possible unvoiced" segment has been found, a new parameter ρ that measures the degree of unvoicing is computed.. The highest it is, the highest is the probability for the segment to be unvoiced. A "possible unvoiced" segment is confirmed to be unvoiced when ρ>Sρ, with Sρ=0.6. If ρ≤Sρ , the segment is considered to be voiced with a pitch period T equal to t/n where n and t are unknown. Therefore, the PDA defines three segment categories: "Voiced", "Unvoiced" and "Voiced with unknown pitch".

In summary, the PDA requires three parameters to be optimised which defaults values are: S=0.3, $S_{pe}$=0.5 and $S_\rho$=0.6. In practice, the optimisation of S is sufficient and we performed all experiments with $S_{pe}$=0.5 and $S_\rho$=0.6.

## 4. THE SPEECH DATABASE

### 4.1. The telephone speech

#### *4.1.1. The clean speech*

Forty eight speakers have been randomly selected from a database of 393 speakers collected by the Center for Information Technology Innovation (CITI) (ex. CWARC). Each speaker has pronounced isolated digits and isolated commands from home through the telephone network. The speech was automatically sampled to 8kHz. The duration of the tested speech is of 45 seconds. Twenty-four French words were used, each one pronounced by a female and a male (24 males, 24 females) speaker.

Four French sentences spoken by two females and one male have also been used. The total speech duration of those sentences is equal to 6.7 seconds.

#### *4.1.2. The noisy speech*

A white Gaussian noise has been added to each word or sentence with an averaged signal to noise ratio of 0 dB. A noise generator has been used for each of the speech files.

### 4.2. The vehicle speech

The PDAs have been tested on five speakers (3 males and 2 females). Each one pronounced five French names in various conditions (0 km/h, 60 km/h, 90 km/h and 130 km/h). The duration is approximately 105 seconds.

### 4.3. The clean and laryngograph labelled speech

We used a database kindly provided by P. Bagshaw [1]. It contains approximately 5 minutes of fifty English sentences read by one male and one female.

## 5. EXPERIMENTS AND RESULTS

### 5.1. Pitch deviation criteria and error estimations

For each frame (10ms frame shift), a pitch deviation is estimated. A gross error deviation is observed when the pitch deviation is greater than 20%, otherwise it is a fine error. Let $V_r$ to be the number of voiced reference frames and $UV_r$ the number of unvoiced reference frames. Six evaluation measures are defined as :

$$V/VunP = \frac{\text{nb of } V_r \text{ declared voiced with unknown pitch } (VunP)}{\text{nb of } V_r};$$

$$V/UV = \frac{\text{nb of } V_r \text{ declared unvoiced}}{\text{nb of } V_r};$$

$$UV/V = \frac{\text{nb of } UV_r \text{ declared voiced or VunP}}{\text{nb of } UV_r};$$

$$UV/VunP = \frac{\text{nb of } UV_r \text{ declared voiced with unknown pitch}}{\text{nb of } UV_r};$$

$$\text{Gross error} = \frac{\text{nb of } V_r \text{ declared voiced with } \Delta f > 20\% \text{ or VunP}}{\text{nb of } V_r \text{ declared voiced}};$$

$$\text{Aver}(\Delta f) = \frac{\sum \Delta f \text{ for the nb of } V_r \text{ declared voiced with } \Delta f \le 20\%}{\text{nb of } V_r \text{ declared voiced with } \Delta f \le 20\%}.$$

Detailed results are presented in next sections. The UV/V column includes the UV/VunP plus the UV/V frames with estimated pitch. The gross error includes the number of frames declared voiced with unknown pitch (V/VunP).

### 5.2.Telephone speech

#### *5.2.1. Noisy telephone speech results*

| | V/UV (%) | UV/V (%) | Gr. err. (%) | Aver(Δf) | UV/VunP (%) | V/VunP (%) |
|---|---|---|---|---|---|---|
| AMPEX $\delta_V$=1.2 $\delta_T$=0.05 | 27.2 | 3.6 | 5.6 | 2.2 | Nil | Nil |
| Proposed PDA S=0.3 | 23.94 | 3.04 | 7.25 | 1.76 | 0.08 | 3.21 |

#### *5.2.2. Clean telephone speech results*

| | V/UV (%) | UV/V (%) | Gr. err. (%) | Aver(Δf) | UV/VunP (%) | V/VunP (%) |
|---|---|---|---|---|---|---|
| AMPEX $\delta_V$ = 1.6 $\delta_T$=0.05 | 6.0 | 9.1 | 1.7 | 1.91 | Nil | Nil |
| Proposed PDA S=0.4 no select. | 8.6 | 10.21 | 2.76 | 1.6 | 1.87 | 1.4 |

The proposed PDA has been tested without the selection module (no select) and with automatic channel selection. When the selection module is not active, all channels are used (M=20 in equation (4)). The channel selection was efficient for noisy telephone speech but was not useful for "clean" telephone speech. The proposed PDA yielded better results for noisy data and worst results for the clean telephone speech in comparison to AMPEX.

### 5.3. Car speech

#### *5.3.1. 60 km/h*

| | V/UV (%) | UV/V (%) | Gr. err. (%) | Aver(Δf) | UV/VunP (%) | V/VunP (%) |
|---|---|---|---|---|---|---|
| AMPEX $\delta_V$=1.4 $\delta_T$=0.05 | 13.22 | 9.63 | 13.11 | 3.08 | Nil | Nil |
| Proposed PDA S=0.3 no select. | 11.59 | 8.89 | 3.11 | 2.52 | 1.81 | 1.72 |

#### *5.3.2. 0, 90 and 130 km/h*

| | V/UV (%) | UV/V (%) | Gr. err. (%) | Aver(Δf) | UV/VunP (%) | V/VunP (%) |
|---|---|---|---|---|---|---|
| AMPEX $\delta_V$=1.4 $\delta_T$=0.05 | 15.50 | 7.48 | 6.11 | 3.32 | Nil | Nil |
| Proposed PDA S=0.3 no select. | 12.21 | 7.04 | 3.94 | 2.5 | 1.56 | 1.85 |

The channel selection did not improve the performance and the proposed PDA yielded better results than AMPEX. The gross error of AMPEX is very important for the 60km/h driving conditions.

### 5.4. Clean speech

#### *5.4.1. Male speaker*

| V/UV (%) | UV/V (%) | Gr. err. (%) | Aver(Δf) | UV/VunP (%) | V/VunP (%) |
|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| AMPEX $\delta_V$=1.4; $\delta_T$=0.1 | 5.8 | 4.85 | 2.35 | 3.34 | Nil | Nil |
| Proposed PDA S=0.3 no select. | 15.65 | 2.8 | 1.9 | 1.6 | 0.98 | 1.43 |

*5.4.2. Female speaker*

| | V/UV (%) | UV/V (%) | Gr. err. (%) | Aver(Δf) | UV/VunP (%) | V/VunP (%) |
|---|---|---|---|---|---|---|
| AMPEX $\delta_V$=1.4; $\delta_T$=0.1 | 2.79 | 4.50 | 3.08 | 2.93 | Nil | Nil |
| Proposed PDA S=0.3 no select. | 4.48 | 5.4 | 1.64 | 2.34 | 1.59 | 0.51 |

AMPEX yielded the best results and the channel selection process was not helpful. The proposed PDA yielded comparable results to AMPEX (slightly poor) on the female data, but the performance was poor on the male speaker.

In comparison to eSRPD [1], AMPEX yielded better performance in terms of V/UV and UV/V errors. The proposed PDA yielded comparable results, in terms of V/UV and UV/V errors (worst for the male and better for the female). It is important to notice that AMPEX and the proposed PDA do not assume any apriori knowledge of the speaker's sex. Therefore, they could not yield a smaller gross error than the eSRPD [1].

## 6. DISCUSSION

The proposed PDA (with and without channel selection) outperforms AMPEX on 0dB telephone and car speech. On clean telephone and clean speech AMPEX was better. The robustness of our PDA resides in the use of the channel selection algorithm and in the pitch extraction module. Good performance of AMPEX on the clean telephone and clean speech seems to be due to the continuity pitch analysis performed by the last module of AMPEX (pitch extraction and voiced/unvoiced decision module).

According to the observations made, the proposed PDA does not estimate the pitch value when the task seems to be too difficult (VunP). But when pitch is estimated, the true gross error is less than the one from AMPEX.

Both AMPEX and our PDA are relatively versatile. In fact, we had to optimise only two parameters ($\delta_T$ and $\delta_V$) for AMPEX and one parameter (S) for PDA. Of course, we could improve the performance of our system by optimising $S_{pe}$ and $S_p$ for each database, which we did not do.

We have proposed a PDA for noisy speech and have compared the performance with an auditory PDA. Our PDA performs relatively well when the speech is noisy. Comparisons on clean speech have also been performed and show performance less good than AMPEX, but comparable to those obtained with the eSRPD system, without any assumptions about the sex pitch range. The tests have been performed on fifty eight speakers.

We plan to test the system on a larger clean speech database and in other environments (factory) to take into account the Lombard effect (as is done for the car speech).

## 7. ACKNOWLEDGEMENTS

Thanks are due to Dr. L. Van Immerseel and Dr. J. P. Martens for providing us the original AMPEX software. We also thank Dr. P. Bagshaw for providing us with the clean speech database and the laryngograph contour. Many thanks to the CITI and ALCATEL M.T. for the noisy speech databases. This work has been supported by the NSERC of Canada, by the FCAR of Québec, by the Canadian Microelectronics Corporation and by the "fondation" from Université du Québec à Chicoutimi.

## 8. REFERENCES

[1] P. C. Bagshaw, S. M. Hiller and M. A. Jack (1993), "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching", *Proc. 3 rd european Conference on Speech Communication and Technology, Berlin, 21-23 september*, pp.1003-1006.

[2] B. Delgutte (1980), "Representation of speech-like sounds in the discharge patterns of auditory nerve fibers", *J. Acoust. Soc. Amer.* Vol. 68 (3), pp.843-857.

[3] J. F. Kaiser (1990), "On a simple algorithm to calculate the 'energy' of a signal", *Proc. of IEEE-ICASSP'90, Albuquerque*, pp.381-384.

[4] J. F. Kaiser (1993), "Some Useful properties of Teager's energy operators", *Proc. IEEE-ICASSP, April 93, vol. 3*, pp.149 -152.

[5] Y. Medan, E. Yair and D. Chazan (1991), "Super Resolution Pitch Determination of Speech Signals", *IEEE Trans. on Signal Processing*, Vol. 39, No 1, pp.40-48.

[6] M. I. Miller and M. B. Sachs (1984), "Representation of voice pitch in discharge patterns of auditory nerve fibers", *Hearing Research*, Vol. 14, pp.257-279.

[7] R.D. Patterson (1986), "Spiral Detection of Periodicity and the Spiral Form of Musical Scales", *Journal of Psychology of Music*, Vol. 14, pp.44-61.

[8] J. Rouat, S. Lemieux and A. Migneault (1992), "A spectro-temporal analysis of speech based on nonlinear operators", *Proc. Int. Conf. on Spoken Language Processing, Banff*, Vol. 2, pp.1629-1632.

[9] L. M. Van Immerseel and J-P. Martens (1992), "Pitch and voiced/unvoiced determination with an auditory model.", *J. Acoust. Soc. Amer.* Vol. 91(6), pp.3511-3526.