A pitch determination and voiced/unvoiced decision algorithm for noisy speech

Jean Rouat [*], Yong Chun Liu, Daniel Morissette

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

Abstract

The design of a pitch tracking system for noisy speech is a challenging and yet unsolved issue due to the association of "traditional" pitch determination problems with those of noise processing. We have developed a multi-channel pitch determination algorithm (PDA) that has been tested on three speech databases (0 dB SNR telephone speech, speech recorded in a car and clean speech) involving fifty-eight speakers. Our system has been compared to a multi-channel PDA based on auditory modelling (AMPEX), to hand-labelled and to laryngograph pitch contours. Our PDA is comprised of 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. Our PDA excelled in performance over the reference system on 0 dB telephone and car speech. The automatic selection of channels was effective on the very noisy telephone speech (0 dB) but performed less significantly on car speech where the robustness of the system is mainly due to the pitch extraction module in comparison to AMPEX. This paper reports in details the voiced/unvoiced, unvoiced/voiced performance and pitch estimation errors for the proposed PDA and the reference system while utilising three speech databases.

Zusammenfassung

Der Entwurf eines Systems zur Grundfrequenzanalyse von verrauschter Sprache ist eine anspruchsvolle und bisher noch nicht zufriedenstellend gelöste Aufgabe, da hierbei "traditionelle" Probleme bei der Grundfrequenzextraktion mit Problemen bei der Verarbeitung verrauschter Signale zusammentreffen. Wir stellen einen Mehrkanal-Grundfrequenzalgorithmus (PDA) vor, der mit drei Sprachdatensammlungen mit ingesamt 58 Sprechern getestet worden ist (Telefonsprache mit 0 dB SNR, Sprachsignale, die im Auto aufgezeichnet wurden, sowie unverrauschte Sprachsignale). Das System wurde verglichen mit dem AMPEX System so wie mit manuell erstellten Referenzkonturen und Grundfrequenzkonturen, welche aufgrund des Laryngosignals erstellt wurden. AMPEX ist ein Mehrkanal-PDA, der auf einem Modell des menschlichen Gehörs beruht. Unser PDA besteht aus einem Modul zur automatischen Kanalauswahl und einem Grundfrequenzextraktionsmodul, das zur Extraktion ein pseudoperiodisches Histogramm benutzt (Kombination der normalisierten Skalarprodukte der ausgewählten Kanäle). Das System erwies sich gegenüber dem Referenzsystem bei den 0 dB Telefonsignalen und bei den im Auto aufgenommenen Signalen überlegen. Bei den stark verrauschten Telefonsignalen (0 dB) führte die automatische Kanalauswahl

---

[*] Corresponding author. E-mail: jrouat@uqac.uquebec.ca.

zur Verbesserung, während bei den im Auto aufgezeichneten Signalen die Robustheit des Gesamtsystems hauptsächlich auf ein – im Vergleich zum AMPEX-System – besseres Verhalten des Grundfrequenzextraktionsmoduls zurückzuführen ist. Ausführlich geht der Artikel ein auf die Performanz des Systems und des Referenzsystems für die drei Sprachsammlungen in bezug auf Stimmhaft/Stimmlos-Fehler, Stimmlos/Stimmhaft-Fehler und Grundfrequenzfehler.

Résumé

La conception d'un système de suivi de fréquence glottale, pour de la parole bruitée, est complexe et constitue un problème qui est loin d'être résolu. En effet, le traitement en milieu bruité est une difficulté supplémentaire qui s'ajoute à celle du suivi de la fréquence glottale. On propose ici un algorithme de détermination de fréquence glottale qui est basé sur une analyse multicanaux. Cet algorithme a été testé sur 3 bases de données (parole téléphonique bruitée artificiellement à 0 dB, enregistrement dans une automobile et parole "propre") regroupant cinquante-huit locuteurs. Le système a été comparé à AMPEX (modèle auditif) et à des contours de fréquence glottale obtenus de façon manuelle ou par laryngogrammes. Notre algorithme inclut un module de sélection automatique des canaux significatifs ainsi qu'un module d'extraction de fréquence glottale basé sur un pseudo-histogramme périodique (obtenu par combinaison de produits scalaires normalisés des signaux provenant des canaux sélectionnés). Sur les enregistrements bruités (voiture et parole téléphonique à 0 dB), le système proposé dépasse AMPEX. Il a été observé que la sélection automatique des canaux améliore les performances sur la parole à 0 dB mais pas sur les enregistrements en véhicule automobile. L'article décrit le système proposé ainsi que les performances en termes de décisions voisé/non voisé, d'erreur fine et grossière.

1. Introduction

The automatic tracking of pitch has multiple applications in the field of speech processing and speech technology. It has been demonstrated that prosody can provide the principal cue for resolving some syntactic ambiguities and studies are being developed to include prosodic information into various continuous speech recognition systems. Pitch contour is also useful in assisting hearing impaired people. Pitch determination might facilitate the diagnosis of aphasia and dysarthria as well as be integrated in computer aided pronunciation teaching systems. One could continue to enumerate many other potential applications based on the automatic determination of pitch. However, most of them are limited to clean speech and cannot be implemented in real life applications, due to the difficulty of determining pitch of degraded speech. Pitch tracking of degraded speech is a challenging and yet unsolved issue that includes the intrinsic problems of "standard" PDA (first formant close to the fundamental, speakers variabilities, ...) in combination with obtaining the information out of the noise or the interference. This paper proposes a new PDA system for noisy speech and evaluates the performance in real and artificial noisy situations.

Reference is made to three distinct PDA categories: time-domain, frequency-domain, and time-frequency domain. Time-domain PDAs determine pitch by essentially relying on zero-crossings, envelopes, peak positions, cross-correlations of speech signals or band-limited signals and comb filtering to measure the averaged pitch value and/or to locate the instants of glottal opening or closure. The frequency-domain PDAs exploit the harmonic structure of the short-term speech spectrum in order to identify the fundamental frequency. A majority of time-frequency domain PDAs perform a time domain analysis of band-limited signals obtained via a multi-channel system.

Among the recent pitch determination algorithms one can refer to (Van Immerseel and Martens, 1992) and (Medan et al., 1991). Herewith, properties such as pitch estimation with good resolution and robustness are reported.

Van Immerseel and Martens (1992) propose a pitch and voiced/unvoiced determination algorithm (AMPEX), based on an auditory model. They compare their PDA with the SHS (Hermes, 1988) and the SIFT (Markel,

1972) PDAs on clean and noisy high-pass-filtered speech. AMPEX is reported to be one of the best evaluated PDA, as the database is comprised of 14 male and 14 female speakers for a total speech duration of 5600 frames (one frame has a duration of 10 ms) covering the entire database.

Medan et al. (1991) propose a time-domain super-resolution pitch algorithm and report that the pitch determination of an extended sentence with white Gaussian noise added in two levels of 10 and 3 dB SNR provides acceptable results.

One should cite the work of Seneff (1985, 1988) who proposed a speech system based on multi-band analysis and synchrony detection. Seneff points out that the GSD ("Generalised Synchrony Detection") is able to enhance the bands for which the output is closely synchronised to the centre frequency of the considered channel. Hunt and Lefèbre (1988) modify the original Seneff model and report speech recognition experiments on corrupted speech. The results suggest that a multi-band analysis with "channel selection" is a promising approach. By channel selection, we mean an automatic process that chooses the less corrupted bands.

A partial list of pitch determination techniques can be found in (Hess, 1983) and (Hermes, 1992).

2. Pitch determination of noisy speech

The speech that we used is spoken in the settings of a car, the telephone network (clean or artificially corrupted with noise) and a quiet room. Telephone speech does not contain strong energy in the low and high frequency components. Speech in a very noisy environment can be degraded in such a way that the low-frequency components become entirely unreliable. Moreover, in real life applications, the PDA has to be insensitive to the quality of the telephone microphones, the various distortions introduced by the telephone system, the car speed, the surface quality of the road, etc.

A second consideration lies in evaluating the pitch determination algorithms. Usually a reference pitch is required to evaluate the performance of a proposed PDA. Hess reports that: "in pitch determination, the bases of comparison are difficult to generate or totally missing. There is no PDA (Pitch Determination Algorithm) that operates without errors, and there is such a variety of signals and marginal conditions that we cannot test any situation in advance. Thus, it becomes understandable that designers of new PDAs have seldom provided detailed data on the performance of their algorithms" (Hess, 1980, Section 5, p. 632). The reference pitch is usually obtained by labelling clean speech, while noise is further added during the experiment.

A third consideration is related to the difficulty of linking results with estimation of the signal to noise ratios. Most of the work reported on PDAs evaluates the performance based on the average signal to noise ratio for a single sentence or for an entire database. Again, it is very difficult to compare the results from two dissimilar PDAs if the speech database is not identical, even if the average signal to noise ratio may be similar. Typically, a PDA is optimised for a specific task (range of noise, range of speakers, etc.) making it extremely difficult to design a PDA with comparable performance on clean and noisy speech without modification of the PDA parameters.

This paper addresses the above considerations and proposes a pitch determination algorithm that is able to track the pitch and is resistant to noise. Three existing databases are used to test the system and to compare with a reference PDA system (AMPEX). Both systems are evaluated with hand-labelled data or laryngograph estimates. Scores are presented for clean and noisy speech.

In the present work we are interested in a PDA suitable for various conditions in comparison with most work presented in the literature where the number of speakers is relatively limited and the conditions are quite specific. The proposed and reference PDAs were tested on 51 speakers (25 males and 26 females) utilising the telephone database. The corpus contained 24 isolated words and four sentences spoken through the telephone network, with different telephone microphones. Gaussian white noise was added to speech and the average signal to noise ratio (for the first database) was approximately 0 dB. The second database, recorded in a car under various driving conditions was comprised of five speakers pronouncing five names. The third database

was comprised of five minutes of clean speech pronounced by a female and a male speaker and labelled with laryngograph data; this database was used to compare the accuracy of the different PDAs. Our PDA was superior to AMPEX relatively to the noisy data (telephone and car).

3. The reference PDA

Comparison of PDAs is far from being trivial since each PDA is different and typically optimised for a specific task. As the speech to noise ratio was small, we were unable to use a reference PDA based on LPC or FFT analysis. The proposed PDA includes a temporal analysis of speech performed by combining "correlation" information from different channels. Therefore we decided to use a similar reference PDA based on a multi-channel and temporal analysis with a combination of correlation functions.

3.1. The preprocessor

AMPEX (Auditory Model-based Pitch Extractor: Van Immerseel and Martens, 1992; Van Immerseel, 1993) is a pitch extractor based on an auditory model (Martens and Van Immerseel, 1990) designed for envelope analysis in order to explain human auditory thresholds of multitone masking and modulation (Martens, 1982). The model comprises of a middle ear filter, a bank of auditory filters, a model of the mechanical to neural transduction and auditory nerve transmission, plus a model of virtual tone component: e_v for each channel. The PDA estimates the pitch based on the virtual tone component by performing a correlation analysis of the e_v for each channel.

AMPEX searches for pitch frequencies within the range 62.5–500 Hz. Van Immerseel and Martens propose a fast representation and implementation of the correlation function (pseudo-autocorrelation). A global pseudo-autocorrelation function $R(m)$ is obtained by accumulating the pseudo-autocorrelation across the channels and is then smoothed.

3.2. The pitch extraction and the voiced / unvoiced decision

AMPEX searches for peaks in $R(m)$, and for each peak that exceeds a threshold δ_r, a pitch candidate and an evidence factor are generated.

The final pitch $T_0(n)$ and its evidence $E_0(n)$ for frame n are derived from the preliminary candidates generated for frames $n - 2$ to $n + 2$. The pitch for frame n is always selected from the set of candidates for that particular frame. The pitches and evidences from the other frames $(n - 2, n - 1, n + 1, n + 2)$ are used to weight the candidates from frame n when a continuity between the candidates from frame n and the considered pitch candidates from the frames $(n - 2, n - 1, n + 1, n + 2)$ is found. If no candidate prevails for T_0, then the pitch and its evidence are set to zero. There is a continuity between two pitch candidates, T_1 and T_2, if

$$\left| \frac{T_1 - T_2}{T_1 + T_2} \right| \leqslant \delta_T, \tag{1}$$

where δ_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 the evidence is greater than a fixed threshold δ_v, or greater than $\delta_v/2$ with a continuity between $T_0(n)$ and $T_0(n - 1)$.

According to Van Immerseel and Martens there are two parameters that need to be optimised: δ_T and δ_v; δ_r is not crucial.

3.3. Modifications

The original AMPEX system has been adapted for this study from 10 kHz sampling rate to 8 kHz (Van Immerseel). We modified the original threshold parameters to improve the AMPEX performance on our databases.

4. 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. Licklider (1956) showed that the low pitch of the residue persists even when the low-frequency channels are masked with low-frequency noise. Terhardt proposes a procedure of "virtual-pitch" perception demonstrating that the pitch can be obtained even when the fundamental is absent from the signal (Terhardt, 1979; Terhardt et al., 1982). The work by Delgutte (1980) confirms that some auditory fibers show modulation patterns corresponding to harmonic interactions (beats). Therefore, the auditory system is able to track pitch by relying on patterns of modulation for fibers influenced by a summation of stimulus harmonics (Delgutte, 1980; Miller and Sachs, 1984). Such modulations are observed at the output of a perceptive filter-bank (Patterson, 1986; Rouat et al., 1992) where beats of harmonics are observed in medium and high-frequency channels (channels where harmonics are unresolved). Our model relies on a combination of the modulation information (beats) and on the resolved harmonics (low frequency channels) to determine the pitch that is available through the channels.

The proposed PDA comprises three modules (Fig. 1). The first module is an auditory filter bank. The second module processes the output of each filter to enhance the modulation periodicity from the signal and combines the incoming information from the channels into a "pseudo-periodic histogram". We refer to this as the
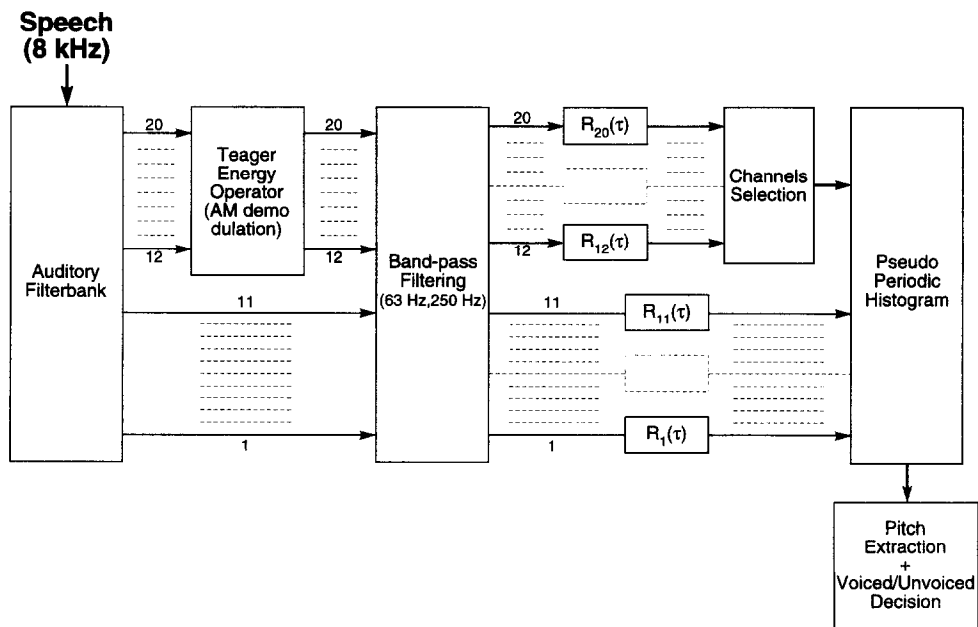

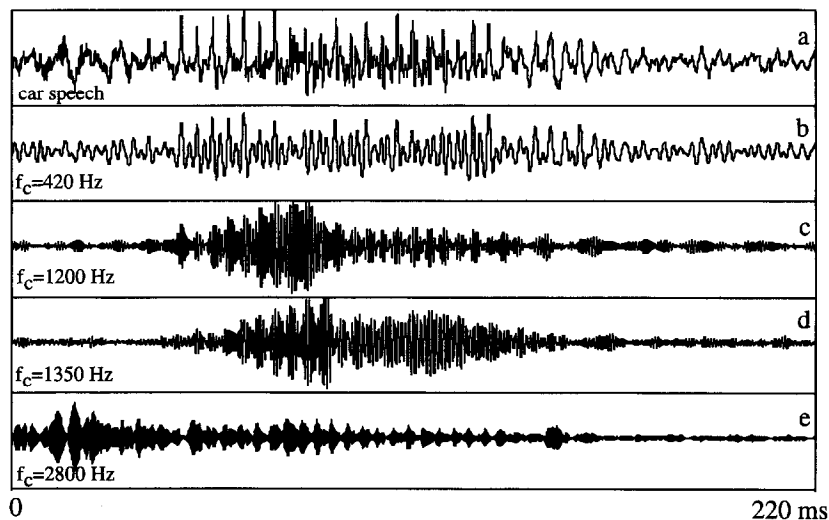Fig. 1. The proposed pitch detemination algorithm.

Fig. 2. Cochlear filter output for the French name "Jean" spoken in a car by a female speaker. (a) Raw speech; (b) filter 3 (centre frequency: 420 Hz); (c) filter 11 (centre frequency: 1200 Hz); (d) filter 12 (centre frequency: 1350 Hz); (e) filter 18 (centre frequency: 2800 Hz).

"spectro-temporal analysis module". The third module estimates the pitch. The speech has been sampled to 8 kHz after proper low-pass filtering and down-sampling, depending on the database.

4.1. The auditory filterbank

The auditory filterbank simulates twenty groups of inner hair cells and covers the frequency range from 330 Hz to 3700 Hz. It is known that low-frequency auditory nerve fibers tend to phase lock to the stimulus. High-frequency fibers however do not resolve the high frequency components and the temporal features are coded similarly to those of amplitude-modulated tones by extracting the envelope of the band-limited signal coming from the inner hair cells. Therefore, in the PDA that is presented, we consider that phase locking with the stimulus is observed in the low frequency channels (centre frequency for channels 1–11: 330–1270 Hz) and that envelope extraction is performed by the medium and high-frequency channels (centre frequency for channels 12–20: 1400–3700 Hz). Of course, the proposed dichotomy is a crude simplification of what is occurring in the peripheral auditory system.

The first eleven auditory filters have an Equivalent Rectangular Bandwidth (ERB) of one critical band unit; in accordance with the work of Patterson (1976) and Moore and Glasberg (1983) (Glasberg and Moore, 1990). The ERB of an auditory filter is the bandwidth of a rectangular filter that has an identical peak transmission as the auditory filter and delivers similar total power in case of white noise input (Moore, 1989). The last nine filters (number 12–20) are based on the AMPEX system and duplicate the last nine filters of the reference PDA (Van Immerseel, 1993). According to Van Immerseel, the speech is upsampled from 8 kHz to 16 kHz prior to filtering with the last nine AMPEX filters.

The output of four of these filters is shown in Fig. 2. Section a is the original speech signal as processed by the system (the French name "Jean" spoken by a female speaker in a car, driving at 60 km/h). Sections (b)–(e) are the output of channel 3 (420 Hz centre frequency), channel 11 (1200 Hz), channel 12 (1350 Hz) and channel 18 (2800 Hz).
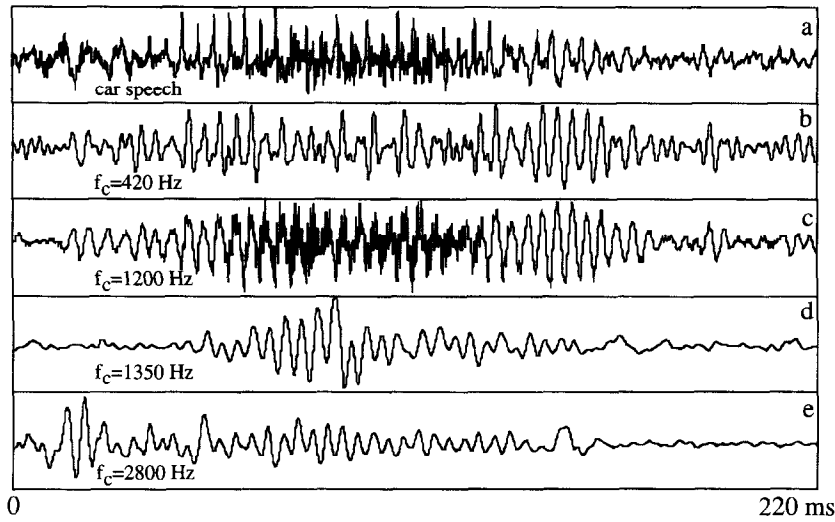
Fig. 3. Intermediate signals before computing the normalised scalar product. (a) Raw speech; (b) band-pass filtered output of channel 3; (c) band-pass filtered output of channel 11; (d) band-pass filtered output of Teager energy operator, channel 12; (e) band-pass filtered output of Teager energy operator, channel 18.

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 precisely imitate the peripheral auditory system.

4.2. The spectro-temporal module

4.2.1. Low-frequency channels

Channels 1–11 and channels 12–20 are subject to different processing. Channels 1–11 are low-pass filtered at 250 Hz using a 3rd order Butterworth filter. Then, the very low frequency distortions (including the DC value) are removed by high-pass filtering to 63 Hz (FIR, window length of 16 ms). Fig. 3(a–c) presents the original signal ("Jean") and the intermediate signals for channels 3 and 11.

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 $i = 1, \ldots, 11$. The first component of vector $X_i(j)$ is the $(j - N/2)$th signal sample and corresponds to the first point in the centred analysis window. $X_i(j)$ and $Y_i(j)$ are $(N + 1)$-dimensional vectors:

$$X_i(j) = \{ x(j - N/2), \ldots, x(j + N/2) \}$$

and

$$Y_i(j) = X_i(j + \tau) = \{ x(j - N/2 + \tau), \ldots, x(j + N/2 + \tau) \}.$$

Therefore,

$$R_i(\tau) = \frac{1}{\sqrt{E_x}} \frac{1}{\sqrt{E_y}} \sum_{n = -N/2}^{N/2} x(j + n) x(j + n + \tau), \quad \tau = 13, \ldots, 2N, \tag{2}$$

with

$$E_x = \sum_{n = -N/2}^{N/2} x^2(j + n) \quad \text{and} \quad E_y = \sum_{n = -N/2}^{M/2} x^2(j + n + \tau).$$
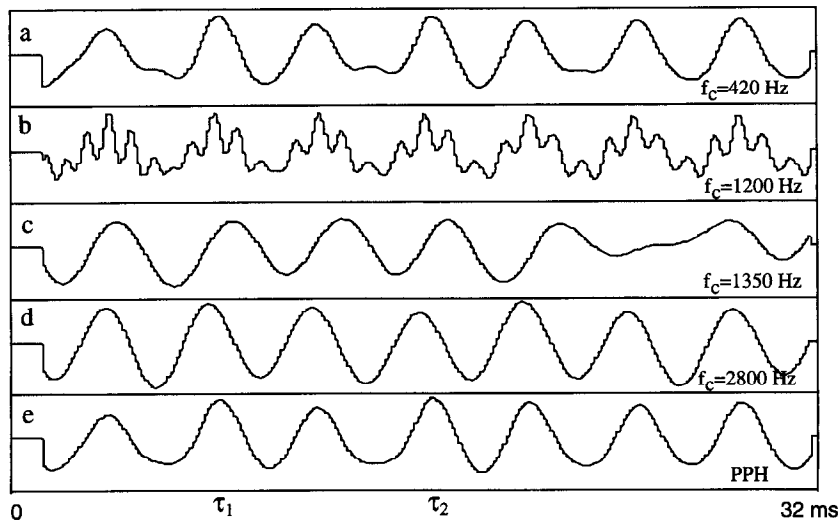
Fig. 4. Normalised scalar product and PPH (frame centred on 88 ms, see also Figs. 2 and 3). (a) Channel 3; (b) channel 11; (c) channel 12; (d) channel 18; (e) combined scalar products (PPH).

E_x is the present segment energy, E_y is the energy of the shifted segment and N is equal to 128. Actually, the scalar product is calculated for $\tau = 13$ to 256, equivalent to search for pitch frequencies within the interval of 62.5–600 Hz at 8 kHz sampling rate with at least two periods observed in $R_i(\tau)$ when τ varies from 0 to 32 ms. $R_i(\tau)$ is reported in Fig. 4(a,b) for channels 3 and 11. The analysis window for that particular frame in the signal is centred on 88 ms.

4.2.2. Medium and high-frequency channels

The output of channels 12–20 is a bandpass signal with a narrow-band spectrum centred on the central frequency (CF) of each channel. The output signal $s(t)$ from a channel can be considered to be an amplitude – and phase – modulated cosine of frequency f_c:

$$s(t) = A(t)\cos\left[2\pi f_c t + \phi(t)\right],\qquad(3)$$

where $A(t)$ is the modulating amplitude and $\phi(t)$ is the modulating phase.

Channels 12–20 are processed using the Teager Energy operator (Kaiser, 1990) in order to approximate the signal envelopes. Kaiser proposes the Teager energy operator E_n capable to extract the energy of a signal based on mechanical and physical considerations. For a digital signal s_n we have

$$E_n = s_n^2 - s_{n+1}s_{n-1}.\qquad(4)$$

This operator is able to track either the envelope of an AM signal or the instantaneous frequency of an FM signal (Maragos et al., 1991). In the present work, we assume that the AM term is dominant in the medium- and high-frequency channels in comparison to the FM term. Therefore, we neglect the FM term that is considered relatively constant in comparison to the AM term. It should be emphasised that the FM term is related to the instantaneous frequency and consequently dominated by the characteristic frequency of each filter (Rouat, 1993). As our algorithm includes high-pass filtering and normalised scalar products, this FM component does not affect our results.

The Teager Energy operator (Kaiser, 1993) output is then smoothed and high-pass filtered by performing similar filtering (low-pass and high-pass filtering) as described in the previous section. Thereafter, the signals

are down sampled from 16 kHz to 8 kHz before computing $R_i(\tau)$ according to Eq. (2). Figs. 3(d,e) and 4(c,d) present the intermediate signals for channels 12 and 18.

4.2.3. Channel selection for medium and high frequencies

The objective of the channel selection module is to maintain the channels that carry relevant pitch information. For each channel i, a new normalised scalar product $R'_i(\tau)$ is evaluated with a larger signal window (30 ms) than the original $R_i(\tau)$ (16 ms). The same time lag interval is used ($\tau = 13$ to 256). If the new scalar product $R'_i(\tau)$ exhibits a local maximum at the same time lag τ_{max} as the absolute maximum of the original $R_i(\tau)$, then the channel is selected. The new product $R'_i(\tau)$ is computed only when the energy of the current envelope, in channel i, exceeds a reference energy or if the previous speech segment has been judged voiced by the Pitch Extraction Module (Section 4.4). The reference energy is the energy of the last nonselected envelope segment for the same channel i. Furthermore, $R'_i(\tau)$ is evaluated only on eleven points centred around the absolute maximum peak in $R_i(\tau)$.

4.2.4. Channel combinations

A new "global" scalar product, referred to as Pseudo-Periodic Histogram (PPH) is derived by the sum of all the $R_i(\tau)$ over all the low-frequency channels and the selected medium/high-frequency channels yielding

$$\mathrm{PPH}(\tau) = \frac{1}{M} \sum_{i=1}^{M} R_i(\tau),$$ (5)

where M is the number of channels that contribute to the pitch. As channels 1–11 are systematically taken into consideration, M is at least equal to 11. The PPH for the particular frame centred on 88 ms is reported in Fig. 4(e).

4.3. Discussion

4.3.1. The first eleven cochlear filters

If $H(f)$ is the transfer function of an ERB filter, f_c the characteristic frequency, g the normalised frequency and p a parameter that controls the passband, then $H(g)$ is given by

$$H(g) = (1 + pg)e^{-pg},$$ (6)

where g is equal to $|f - f_c|/f_c$ and p determines the bandwidth and slopes of the filter (Patterson et al., 1982). From Eq. (6), it can be deduced, that energy will continue to be present remote from the filter characteristic frequency (CF). For example, a filter with a CF around 1000 Hz, will allow the transmission of low-frequency components around 250 Hz with a significant attenuation. The Butterworth low-pass filtering enhances the low-frequency components of the first eleven cochlear filters, yielding eleven representations of the fundamental and low-frequency harmonics.

4.3.2. Medium and high-frequency channels

The nonlinear processing performed by the Teager energy operator allows us to obtain the envelope of the AM signal created by adjacent harmonics ringing in the same cochlear filter. Therefore, the system is able to track the fundamental even if the first partial is absent from the signal.

4.4. The pitch extraction module

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 are fewer than two "valid" peaks in the PPH, the segment is declared unvoiced.

We assume that the two largest peaks correspond to $\tau = \tau_1$ and $\tau = \tau_2$ where τ_1 and τ_2 are multiple or equal to the pitch period T. In other words, T is a common sub-multiple of τ_1 and τ_2 or can be equal to τ_1 or τ_2.

The algorithm searches for the sub-multiples of τ_1 and τ_2 and the pitch period T is the lowest common sub-multiple with PPH(T) being, at least, one of the first ten highest "valid" peaks with

$$PPH(T) \geqslant S_{pe} \max[PPH(\tau_1); PPH(\tau_2)], \tag{7}$$

where $S_{pe} = 0.5$.

If the algorithm fails to find such a T, the segment is declared as being "possibly unvoiced", otherwise it is considered as being "voiced" and the pitch frequency is equal to $1/T$.

4.5. Voiced / unvoiced decision

It is assumed that voiced segments identified by the pitch extraction module are really voiced. Therefore, when a "possibly unvoiced" segment has been identified a new parameter ρ is computed. ρ takes into consideration the dissimilarity between PPH(τ) and its time shifted version PPH($\tau + t_{max}$), where t_{max} locates the highest peak from the first half of PPH(τ) ($\tau = 13, \ldots, N$):

$$\rho = \frac{\max(\text{diffPPH}(\tau)) - \min(\text{diffPPH}(\tau))}{\max(\text{PPH}(\tau)) - \min(\text{PPH}(\tau))} \quad \text{for } \tau = 13, \ldots, N, \tag{8}$$

with diffPPH(τ) = PPH(τ) − PPH($\tau + t_{max}$) and $t_{max} = \text{argmax}_{\tau = 13, \ldots, N}(\text{PPH}(\tau))$.

To a certain extent ρ measures the degree of unvoicing. The higher it is, the greater the probability for the segment to be unvoiced. A "possibly unvoiced" segment is confirmed to be unvoiced when $\rho > S_\rho$, with, $S_\rho = 0.6$. If $\rho \leqslant S_\rho$, the segment is considered to be voiced with unknown pitch. Therefore, the PDA defines three segment categories: "voiced", "unvoiced" and "voiced with unknown pitch".

4.6. Postprocessing

The PDA defines robust islands of voiced speech in order to minimise the incidence of confusion with unvoiced frames. This is particularly important for speech recognition applications, where false alarms can be a problem. A postprocessor is not essential for most applications relying only on voiced islands (speaker identification, speech recognition with specific vocabulary, etc.). We however did use a postprocessor in the study. A robust island of voiced speech is a sequence of at least five "voiced" or "voiced with unknown pitch" frames. The frames can be separated by one unvoiced frame. The postprocessor can be described as follows:

1. As soon as one island is found, a search for the median pitch frequency of the "voiced" frames of this island is performed.
2. For each frame with a pitch frequency greater than 1.9 times the median frequency, the pitch is adjusted by taking the closest sub-multiple of τ_1 and τ_2 (see Section 4.4 for τ_1 and τ_2) which is greater than T. If there is no such sub-multiple, then the frame is classified as "voiced with unknown pitch".
3. For each "voiced with unknown pitch" frame inside the island, a pitch frequency is associated by taking the average of the three closest "voiced frames" from the island.
4. If two islands are separated by an "unvoiced" segment shorter or equal to 40 ms, they are merged and the "unvoiced" segment is labelled as "voiced with unknown pitch" as the system underestimates the number of voiced segments.
5. Finally, all isolated 10 ms "voiced" segments are labelled as "unvoiced".

In summary, the PDA required three parameters to be optimised: S, S_{pe} and S_p. Within the study, it proved sufficient to optimise only S, and we conducted all experiments with the default values $S_{pe} = 0.5$ and $S_p = 0.6$ for the other two parameters. S was optimised by conducting preliminary experiments on a subset (1/3 of the speakers, randomly chosen) with S equal to 0.25, 0.3, 0.35 (original default value), 0.4 or 0.45. The postprocessing uses a buffer length that is dynamically changed to 50 ms + the duration of the island under analysis.

5. The speech database

5.1. The telephone speech

Forty eight speakers were randomly selected from a database of 393 speakers provided by the Center for Information Technology Innovation (CITI). Each speaker was requested to pronounce isolated digits and isolated commands from their home environment to the CITI using the telephone network. The speech was automatically sampled to 8 kHz by the CITI's telephone system with a resolution of 16 bits. We were able to notice a significant difference in the quality of the telephones used by the speakers. The duration of the tested data is approximately 40 seconds. The vocabulary includes the first ten numerical digits in French, 14 command words ("annulation", "Anglais", "Français", "oui", "non", "recommencer", etc.) pronounced by 24 female speakers and 24 male speakers. Four brief French sentences spoken by two female and one male speaker have been extracted from another clean speech database and low-pass filtered to 3300 Hz prior to being sampled at 8 kHz. The data were then high-pass filtered to partially simulate the low-frequency filtering of the telephone system, by a 10th-order FIR filter with a cut-off frequency of 220 Hz. The total duration of the four sentences equals 6 seconds.

5.1.1. The reference pitch

The reference pitch contour was estimated manually from clean speech. The same individual performed the labelling using the Signalyze software package. The reference pitch was estimated frame by frame with a window length of 20 ms. The pitch equals the averaged value of the glottal time intervals occurring in the frame. When unvoiced and voiced speech was present in the same frame, the speech was labelled as voiced when more than half of the frame was voiced.

5.1.2. The noisy speech

White Gaussian noise was added to each word or sentence with an average signal to noise ratio of 0 dB. A noise generator was used for each of the speech files. Consequently, a different white Gaussian noise was added. It is important to note that the level of white Gaussian noise has been approximately estimated for an SNR of 0 dB on a signal where speech is present. In other words, SNRs of $-\infty$ are not taken into account when estimating the average SNR of the database.

5.2. The vehicle speech

The vehicle speech was recorded in a Peugeot 405 GR. The microphone was located on the top left frame of the front windshield and the speech was sampled to 16 kHz with a resolution of 16 bits. The signal was down-sampled to 8 kHz after proper low-pass filtering. The PDAs were tested on five speakers (3 male and 2 female). Each pronounced five French names under various speed conditions (0 km/h, 60 km/h, 90 km/h and 130 km/h). The duration is approximately 105 seconds.

Before testing and labelling, the database was high-pass filtered to 125 Hz. In fact, without filtering, it was not possible to distinguish, by visual inspection, the speech from noise in a significant portion of the data. The

labelling presented difficulties in particular for the $60\,\text{km/h}$ conditions. 5% of the database could not be hand-labelled, even with implementation of various filtering techniques, due to the distortions. Therefore, these files have not been included in the evaluation procedures. When unvoiced and voiced speech was present in the same frame, the speech was also labelled as voiced when more than half of the frame was voiced. A reference pitch is generated at time $(T_1 + T_2)/2$ with T_1 and T_2 corresponding to adjacent glottal pulses (estimated on the speech signal). The duration $T_2 - T_1$ is converted to Hertz and is used as a reference pitch frequency contour.

5.3. The clean and laryngograph labelled speech data

The database (Bagshaw et al., 1993) contains approximately 5 minutes of fifty English sentences read by one male and one female speaker. Also included were utterances containing voiced fricatives, nasals, liquids and glides. The speech was originally sampled to $20\,\text{kHz}$ with a 16-bit A/D converter and was down-sampled to $8\,\text{kHz}$ for the current study.

The reference contour was developed from laryngograms recorded simultaneously with speech by using a pulse location algorithm (Bagshaw et al., 1993). The laryngeal frequency at time $(T_1 + T_2)/2$ is equal to $1/(T_2 - T_1)\,\text{Hz}$ and is used as reference contour in the same manner as for car speech.

6. Experiments and results

This section describes three sets of experiments. The first set has been conducted on clean and noisy telephone speech, the second on five speakers while driving a car in real conditions. The third set of experiments consists of a comparison of the PDAs output with laryngograph data of clean speech.

6.1. Pitch deviation criteria and error estimations

For each frame ($10\,\text{ms}$ frame shift), a pitch deviation is computed according to the relation

$$\Delta f = \frac{|\text{PDA}_{\text{output}} - f_0|}{f_0} \times 100\%, \tag{9}$$

where $\text{PDA}_{\text{output}}$ is the pitch frequency estimated by the tested PDA and f_0 is the pitch frequency associated with the hand-labelled or laryngeal data. Let V_r to be the number of voiced reference frames. We define the averaged fine error $\text{Aver}(\Delta f)$ as

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

Let N be the total number of frames processed by the system:

$$N = N_{\text{V}/\text{VunP}} + N_{\text{V}/\text{UV}} + N_{\text{UV}/\text{V}} + N_{\text{UV}/\text{VunP}} + N_{\text{Gross}} + N_{\text{Corr}}, \tag{10}$$

with $N_{\text{V}/\text{VunP}}$ = voiced frames classified as voiced with unknown pitch, $N_{\text{V}/\text{UV}}$ = voiced frames classified as unvoiced, $N_{\text{UV}/\text{V}}$ = unvoiced frames classified as voiced, $N_{\text{UV}/\text{VunP}}$ = unvoiced frames classified as voiced with unknown pitch, N_{Gross} = voiced frames classified as voiced but with a Δf greater than 20%, N_{Corr} = voiced frames correctly classified with a Δf less than or equal to 20% or unvoiced frames also correctly classified. The value of 20% has been used to ease the comparison with AMPEX.

We define five error measures:

$$\epsilon_{\text{V}/\text{UV}} = N_{\text{V}/\text{UV}}/N, \qquad \epsilon_{\text{UV}/\text{V}} = N_{\text{UV}/\text{V}}/N, \qquad \epsilon_{\text{Gross}} = N_{\text{Gross}}/N,$$

$$\epsilon_{\text{UV}/\text{VunP}} = N_{\text{UV}/\text{VunP}}/N \quad \text{and} \quad \epsilon_{\text{V}/\text{VunP}} = N_{\text{V}/\text{VunP}}/N.$$

Table 1
Telephone speech
(a) Noisy telephone speech results

	$\epsilon_{V/UV}$ (%)	$\epsilon_{UV/V}$ (%)	ϵ_{Gross} (%)	Aver(Δf)	$\epsilon_{UV/VunP}$ (%)	$\epsilon_{V/VunP}$ (%)
AMPEX $\delta_v = 1.2$, $\delta_T = 0.05$	16.32	1.44	2.03	2.2	nil	nil
Proposed PDA $S = 0.3$	14.38	1.18	3.17	1.76	0.03	1.93
Proposed PDA $S = 0.25$, no select.	15.03	1.37	3.76	1.78	0.03	1.84

(b) Clean telephone speech results

	$\epsilon_{V/UV}$ (%)	$\epsilon_{UV/V}$ (%)	ϵ_{Gross} (%)	Aver(Δf)	$\epsilon_{UV/VunP}$ (%)	$\epsilon_{V/VunP}$ (%)
AMPEX $\delta_v = 1.6$, $\delta_T = 0.05$	3.6	3.64	0.75	1.91	nil	nil
Proposed PDA $S = 0.45$	4.35	4.60	0.87	1.61	1.05	0.62
Proposed PDA $S = 0.4$, no select.	5.16	3.33	1.49	1.6	0.75	0.84

δ_v, δ_T and S are thresholds described in Sections 3.2 and 4.4. For a full explanation of the other symbols, please see Section 6.1.

Depending on the application, the reader may be interested in fine or gross error estimation instead of V/UV and UV/V errors or vice versa. Therefore, we present the detailed results in Tables 1–3, and we comment on "total gross error" in the text. The "total gross error" is defined as

$$\epsilon_{tot} = \epsilon_{V/UV} + \epsilon_{UV/V} + \epsilon_{UV/VunP} + \epsilon_{Gross}.$$

The $\epsilon_{V/VunP}$ error could be included in the "total gross error". Once again, depending on the application, the reader can consider that the number $N_{V/VunP}$ defines a gross error in pitch estimation and should be included in the expression of ϵ_{Gross}. On the contrary, other readers can consider that $N_{V/VunP}$ is not an error as the voiced decision has been taken correctly. For the purpose of this paper we do not include the number of voiced frames declared "voiced with unknown pitch" into the total gross error but we provide the detailed results in Tables 1–3.

6.2. Telephone speech

Table 1 reports results for noisy and clean telephone speech. The number of speech frames processed by the systems is equal to 3210 (60% are voiced, 40% unvoiced). The proposed PDA was tested without the selection

Table 2
Car speech
(a) 60 km/h

	$\epsilon_{V/UV}$ (%)	$\epsilon_{UV/V}$ (%)	ϵ_{Gross} (%)	Aver(Δf)	$\epsilon_{UV/VunP}$ (%)	$\epsilon_{V/VunP}$ (%)
AMPEX $\delta_v = 1.4$, $\delta_T = 0.05$	5.55	5.58	7.89	3.08	nil	nil
Proposed PDA $S = 0.35$	3.52	5.2	2.4	2.63	1.97	1.08
Proposed PDA $S = 0.3$, no select.	4.87	4.1	1.89	2.52	1.04	0.72

(b) 0, 90 and 130 km/h

	$\epsilon_{V/UV}$ (%)	$\epsilon_{UV/V}$ (%)	ϵ_{Gross} (%)	Aver(Δf)	$\epsilon_{UV/VunP}$ (%)	$\epsilon_{V/VunP}$ (%)
AMPEX $\delta_v = 1.4$ $\delta_T = 0.05$	7.3	3.95	4.05	3.32	nil	nil
Proposed PDA $S = 0.35$	5.42	3.39	2.97	2.54	0.96	0.76
Proposed PDA $S = 0.3$, no select.	5.82	2.86	2.67	2.5	0.81	0.88

δ_v, δ_T and S are thresholds described in Sections 3.2 and 4.4. For a full explanation of the other symbols, please see Section 6.1.

Table 3
Continuous clean speech results

(a) Male speaker

	$\epsilon_{V/UV}$ (%)	$\epsilon_{UV/V}$ (%)	ϵ_{Gross} (%)	Aver(Δf)	$\epsilon_{UV/VunP}$ (%)	$\epsilon_{V/VunP}$ (%)
AMPEX $\delta_v = 1.4$, $\delta_T = 0.1$	2.41	2.83	1.08	3.34	nil	nil
Proposed PDA $S = 0.35$	5.87	1.6	1.01	1.62	0.93	0.69
Proposed PDA $S = 0.3$, no select.	6.52	1.06	0.78	1.6	0.56	0.59

(b) Female speaker

	$\epsilon_{V/UV}$ (%)	$\epsilon_{UV/V}$ (%)	ϵ_{Gross} (%)	Aver(Δf)	$\epsilon_{UV/VunP}$ (%)	$\epsilon_{V/VunP}$ (%)
AMPEX $\delta_v = 1.4$, $\delta_T = 0.1$	1.01	2.86	2.72	2.93	nil	nil
Proposed PDA $S = 0.35$	1.03	3.38	1.39	2.38	1.43	0.2
Proposed PDA $S = 0.3$, no select.	1.64	2.42	1.42	2.34	1.0	0.18

δ_v, δ_T and S are thresholds described in Sections 3.2 and 4.4. For a full explanation of the other symbols, please see Section 6.1.

module (no select.) and with automatic channel selection. When the selection module is not active, all channels are used ($M = 20$ in Eq. (5)). The channel selection was efficient for the noisy telephone speech but to a lesser extent for "clean" telephone speech. The proposed PDA yielded better V/UV and V/UV decisions for noisy data and lesser results for clean telephone speech in comparison to AMPEX. Fig. 5 shows the results for noisy speech via a telephone system pronounced by a male speaker.

6.3. Car speech

In Table 2(a) the performance for the 60 km/h condition (city roads) is shown. Table 2(b) summarises the results for 0 km/h, 90 km/h and 130 km/h (rural roads). The channel selection module degrades the performance and both proposed PDAs yielded better results than AMPEX. The gross error of AMPEX is very important for the 60 km/h driving conditions. Fig. 6 shows the results of the French name "Jean-Pierre Rivière" spoken by a female speaker during the 60 km/h condition. Preliminary experiments have shown that the 60 km/h is the worst condition (noisy city, gear shift, etc.) for AMPEX. Surprisingly, the 90 km/h and 130 km/h yielded comparable results to the 0 km/h condition. The total number of processed frames is equal to 2500 for the 60 km/h (42% voiced, 58% unvoiced) and 7730 for the 0, 90 and 130 km/h conditions (47%, 53%).
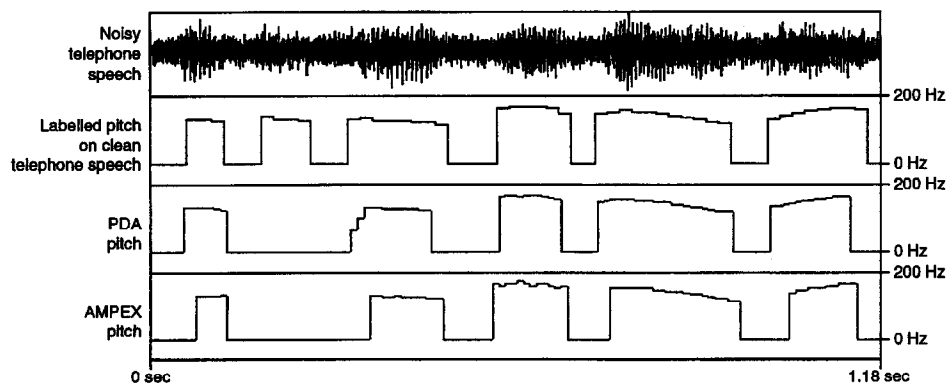

Fig. 5. PDA and AMPEX output for the noisy French telephone sentence "Quand ils ont vu un voyageur" spoken by a male speaker.
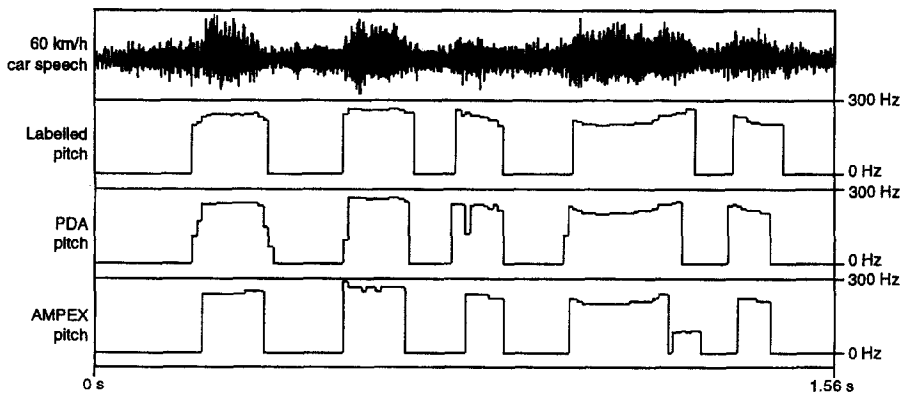
Fig. 6. PDA and AMPEX output for the French name: "Jean-Pierre Rivière" (60 km/h condition, female speaker).

6.4. Clean speech

Both proposed PDAs yielded comparable results to AMPEX (slightly inferior) on the female data, and poor performance on the male speaker. The channel selection process did not improve the results. In comparison with AMPEX, the most frequent errors of the PDA were observed at boundaries between silence or unvoiced frames and voiced segments where the signal is voiced but with irregular PPH. Since the speech of the female speaker included less irregular voiced segments than the male speaker, the performance on the female data was better. It is important to note that AMPEX and the proposed PDA do not assume any a priori knowledge of the speaker's sex. The number of processed frames is equal to 14650 (42% voiced, 58% unvoiced) for the male speaker and 18490 (36%, 64%) for the female speaker.

7. Discussion

Performance of the proposed PDAs (with and without channel selection) was superior to that of AMPEX on noisy telephone and car speech. On clean telephone and clean speech AMPEX performed better. However, in reference to clean speech, on the female speaker, the difference is not overly significant (ϵ_{tot} = 6.59 for AMPEX and ϵ_{tot} = 6.48 for the proposed PDA). On the noisy telephone speech ϵ_{tot} is equal to 19.79 for AMPEX and to 18.76 for the PDA. The voiced/unvoiced and unvoiced/voiced decisions are enhanced with the proposed PDA, however the gross error (3.17%) is greater than with AMPEX (2.03%).

The robustness of our PDA resides in the implementation of the channel selection algorithm and in the pitch extraction module. The channel selection principle is useful on noisy telephone speech where the signal-to-noise ratio (SNR) is very low (0 dB). On car speech, where the signal is distorted, but with reasonable SNR, the selection is not effective and robustness is mainly due to the pitch extraction module. The inferior results of the PDA on clean and clean telephone speech can be explained by the fact that the PDA assumes that voiced frames have sufficiently regular PPH's.

According to observations made, our PDA does not estimate the pitch value when the task seems to be complex (VunP). However, when pitch is estimated, the true gross error is less than for AMPEX with the exception of telephone speech.

The pitch extraction module locates the voiced frames with reliable pitch period values and rejects the segments with irregular PPHs (very noisy frames, vocal fry, etc.). The island postprocessor and the ρ parameter (Eq. (8)) are used to cluster the rejected frames into "unvoiced", "voiced with unknown pitch" and "voiced". If the user plans to run the PDA on clean speech and cannot tolerate the underestimation of the number of voiced frames, the parameter ρ should be optimised for that particular application and the postprocessor should

be used. From our experience it appears that the postprocessor is not crucial when processing very noisy speech and looking only at robust voiced segments.

As stated previously, the design of a PDA suitable for various speech and noise conditions is far from being trivial. In the majority of instances, the PDA has to be adapted to the speech database by optimising various parameters. We found the proposed PDA to be relatively versatile. In fact, we had to optimise two parameters (δ_T and δ_v) for AMPEX and one parameter (S) for our PDA. Furthermore, the same parameter values have been used on the car and clean speech for our PDA in opposition to AMPEX. Of course, we could have improved on the performance of our system by optimising S_{pe} and S_p for each database. In a previous study we identified that a channel selection algorithm, combined with a zero-crossing calculation of correlation functions at the output of an auditory filter-bank yielded better results than those presented here with respect to telephone speech (Rouat and Liu, 1992). The drawback, however, was that the optimisation of the parameters was a very difficult and tedious task when altering the noise and telephone speech conditions.

The comparison with a reference pitch based on hand-labelling of clean speech cannot really reflect the PDA's performance on noisy speech. In fact, for such noisy signal, our auditory system is not always able to locate voiced segments with very low energy, which we require the PDA to identify by referring to the hand-labelled clean speech. It is well known that hand-labelling of very noisy speech can be complex and sometimes unsuccessful in generating a reference pitch. A comparison of the PDA's performance by combining hand-labelling of clean and noisy speech might be an interim solution.

The postprocessor and the channel selection algorithm of our PDA could be improved. In fact, the postprocessor can generate half and third fundamental period values when the pitch variation is large in islands. For example, improvement of the channel selection algorithm can be performed by including a modified "Generalised Synchrony Detection" system as proposed by Hunt and Lefèbre (1988).

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 conducted and show performance less significant than the auditory PDA (AMPEX), but comparable to those obtained with the eSRPD system (Bagshaw et al., 1993) on the same clean speech without any assumptions regarding the sex pitch range.

We intend to further study the system using a larger clean speech database and other environments (factory), taking into account the Lombard effect.

Acknowledgements

References

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. 3rd European Conf. on Speech Communication and Technology*, Berlin, 21–23 September 1993, pp. 1003–1006.

B. Delgutte (1980), "Representation of speech-like sounds in the discharge patterns of auditory nerve fibers", *J. Acoust. Soc. Amer.*, Vol. 68, No. 3, pp. 843–857.

B.R. Glasberg and B.C.J. Moore (1990), "Derivation of auditory filter samples from notched-noise data", *Hearing Research*, Vol. 47, pp. 103–138.

D.J. Hermes (1988), " Measurement of pitch by subharmonic summation", *J. Acoust. Soc. Amer.*, Vol. 83, pp. 257–264.

D.J. Hermes (1992), "Pitch analysis", in: M. Cooke, S. Beet and M. Crawford, Eds., *Visual Representations of Speech Signals* (Wiley, London), pp. 3–25.

W.J. Hess (1980), "Pitch determination – An example for the application of signal processing methods in the speech domain", in: M. Kunt and F. de Coulon, Eds., *Signal Processing: Theories and Applications* (North-Holland, Amsterdam), pp. 625–634.

W. Hess (1983), *Pitch Determination of Speech Signals* (Springer, New York).

M.J. Hunt and C. Lefèbre (1988), "Speaker dependent and independent speech recognition experiments with an auditory model", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 215–218.

J.F. Kaiser (1990), "On a simple algorithm to calculate the 'energy' of a signal", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.'90*, Albuquerque, pp. 381–384.

J.F. Kaiser (1993), "Some useful properties of Teager's energy operators", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, April 1993, Vol. 3, pp. 149–152.

J.C.R. Licklider (1956), "Auditory frequency analysis", in: C. Cherry, Ed., *Information Theory* (Academic Press, New York).

P. Maragos, T. Quatieri and J.F. Kaiser (1991), "Speech non-linearities, modulations and energy operators", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.'91*, Toronto, pp. 421–424.

J. Markel (1972), "The SIFT algorithm for fundamental frequency estimation", *IEEE Trans. Audioprocess.*, Vol. 20, pp. 367–377.

J-P. Martens (1982), "A new theory for multitonal masking", *J. Acoust. Soc. Amer.*, Vol. 72, pp. 397–405.

J-P. Martens and L. Van Immerseel (1990), "An auditory model based on the analysis of envelope patterns", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.'90*, pp. 401–404.

Y. Medan, E. Yair and D. Chazan (1991), "Super resolution pitch determination of speech signals", *IEEE Trans. Signal Process.*, Vol. 39, No. 1, pp. 40–48.

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.

B.C.J. Moore (1989), *An Introduction to the Psychology of Hearing* (Academic Press, London).

B.C.J. Moore and B.R. Glasberg (1983), "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", *J. Acoust. Soc. Amer.*, Vol. 74, No. 3, pp. 750–753.

R.D. Patterson (1976), "Auditory filter shapes derived with noise stimuli", *J. Acoust. Soc. Amer.*, Vol. 59, No. 3, pp. 640–654.

R.D. Patterson (1986), "Spiral detection of periodicity and the spiral form of musical scales", *J. Psychology of Music*, Vol. 14, pp. 44–61.

R.D. Patterson, I. Nimmo-Smith, D.L. Weber and R. Milroy (1982), "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold", *J. Acoust. Soc. Amer.*, Vol. 72, No. 6, pp. 1788–1803.

J. Rouat (1993), "Nonlinear operators for speech analysis", in: M. Cooke, S. Beet and M. Crawford, Eds., *Visual Representations of Speech Signals* (Wiley, London), pp. 335–340.

J. Rouat and Y.C. Liu (1992), "A pitch determination algorithm for very noisy telephone speech", *Proc. ETRW on Speech Processing in adverse conditions*, Cannes, France, 10–13 November 1992, pp. 158, 164–166.

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

S. Seneff (1985), Pitch and spectral analysis of speech based on an auditory synchrony model, Ph.D. Thesis, Research Laboratory of Electronics, MIT.

S. Seneff (1988), "A joint synchrony/mean-rate model of auditory speech processing", *J. Phonetics*, Vol. 16, No. 1, pp. 55–76.

E. Terhardt (1979), "Calculating virtual pitch", *Hearing Research*, Vol. 1, No. 2, pp. 155–182.

E. Terhardt, G. Stoll and M. Seewann (1982), "Algorithm for extraction of pitch salience from complex tonal signals", *J. Acoust. Soc. Amer.*, Vol. 71, pp. 679–688.

L.M. Van Immerseel (1993), Een functioneel gehoormodel voor de analyse van spraak by spraakherkenning, Ph.D. Thesis, Gent University, Belgium.

L.M. Van Immerseel and J-P. Martens (1992), "Pitch and voiced/unvoiced determination with an auditory model", *J. Acoust. Soc. Amer.*, Vol. 91, No. 6, pp. 3511–3526.