Summary of the Thesis: Pitch-based Techniques for Robust Speech Recognition

Juan Andrés Morales Cordovilla

Dpto. de Teoría de la Señal Telemática y Comunicaciones

Universidad de Granada

Abstract

This Thesis proposes and carries out a study of different techniques which, in some way, use the pitch (which will be understood as the fundamental frequency of speech) in order to carry out robust ASR (Automatic Speech Recognition) under noise conditions. The Thesis is not concerned with pitch extraction itself, but with the best way of using pitch for robust speech recognition.

We will also carry out a study of the related bibliography and the state of art regarding these pitch-based techniques for robust ASR. Then, we will propose three pitch-based techniques which will be compared to other similar ones. Our three proposals are: application of *asymmetric windows* to the noisy signal autocorrelation which tries to provide a spectrum less sensitive to noise, two estimators, named as *averaging and sifting estimators*, of the autocorrelation of the clean quasi-periodic signal, and a *noise estimation technique* which can deal with non stationary noise by employing pitch information and which is used to estimate the reliability masks required by a marginalization MD (Missing Data) recognizer.

Additionally, we will discuss the performance limits of the pitch-based techniques for robust ASR which employ minimal assumptions about the noise. In order to do so, we will identify the basic robust mechanisms employed by these techniques for recognizing voiced frames, the optimum mechanisms will be identified (by means of some equivalences), and the corresponding limit results will be experimentally obtained by applying MD oracle masks and ideal pitch. One of our conclusions is that our noise estimation technique for MD recognition is close to the limits of the pitch-based techniques for robust ASR, although it would require additional information in order to achieve the performance with MD oracle masks. Finally, we will comment some possibilities (some of them related to speech without pitch) for future research from the ideas developed in this Thesis.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Summary of the Thesis: Pitch-based Techniques for Robust Speech Recognition

1.1. Introduction

1.1.1. Motivations

Importance of pitch in robust speech recognition

Acoustic noise represents one of the major challenges for ASR (Automatic Speech Recognition) systems. Many different approaches have been proposed to deal with this problem in monaural signal [38, 22, 48] and many of them try to employ some kind of noise information to do robust ASR. However, when one wants to deal with all kind of noises it is clear that the most important information to separate noise from speech is just speech information. There exits many cues and informations which help to distinguish speech from noise but at the end the correct choice will depend on what is defined as speech. Speech can be emitted in many different ways which mainly depend on the considered type of the «main source». These ways can be whispering, vocal harmony speech (in music), etc.. In this Thesis it will be considered that speech is emitted in its normal way, with vibration of the vocal folds and with only one pitch at each time instant.

Continuing with the search for the most important cues, this Thesis will particularly consider the signal pitch due to the three following reasons:

1. Many psychoacoustics experiments, such as those shown in [12, 48], reach the conclusion that very often humans use pitch to separate speech from noise.

2. Pitch is a useful information to distinguish different types of speech segments (voiced, unvoiced and silence) and to separate speech and noise signals.

3. Many robust ASR techniques inspired in human recognition, as shown in [48], use pitch.

Robust techniques based on pitch

The comparison of the different ASR techniques based on pitch is not an easy matter because of several reasons:

1. Each author uses a different pitch extractor to evaluate his technique.

2. It is not clear which is the real cause for obtaining different results: different methods applied to voiced and unvoiced sounds, application of additional techniques (such as cepstral normalization, missing data approaches,...), etc.

3. Sometimes it is not clear whether an author is proposing either a new technique for robust ASR based on pitch or a new robust pitch extractor (or both at the same time).

Because of these reasons, we consider it necessary to do a fair comparison of these pitch-based techniques, trying to show the equivalences between some of them and trying to see the limits of pitch-based recognition. Apart from this, we will propose three new pitch-based techniques but without paying attention to the pitch extractor because this is beyond the scope of this Thesis.

1.1.2. Objectives

Taking into account the previous motivations, the main objectives of the Thesis can be summarized as follows:

1. Recognition of monaural speech which is emitted in its normal way (i.e. with pitch) and contaminated with acoustic noise.

2. Development of a comparative study of both classical and pitch-based techniques for robust speech recognition considered as the state of the art.

3. Development and improvement of robust ASR techniques based on pitch, trying to do minimal assumptions about the noise. In order to do so, we will employ other techniques and recognition schemes such as SS (Spectral Subtraction) or MD (Missing Data).

4. We will show the equivalences between some of the different techniques, doing a fair comparison and trying to answer the question of to what extent recognition can be made more robust by means of the pitch.

1.2. Principles of Automatic Speech Recognition

The first chapters are devoted to explaining some important concepts which will be used throughout the Thesis. These concepts refer to: speech, hearing, signal processing, acoustic representations (cochleagram, spectrogram and cepstrogram) and their masks, pitch extractors, and MD (Missing Data) recognizer based on HMM (Hidden Markov Models).

The most important issues described in these chapters are:

- The «main source model» of speech which considers that speech is a main source which is intensity and spectrally modulated and sometimes replaced by short duration noises (unvoiced sounds). The main source can be a noise in the case of whispered speech, but in a normal situation speech will be identified with a voiced sound and, if pitch is known, the rest of the elements of the speech can be also located (unvoiced sounds and silences) as well. This model is a simplified definition of speech which will be considered to develop a VAD.

- The soft mask of a given time-frequency signal representation (i.e. spectrogram or cochleagram) can be estimated through local SNR estimates or through harmonicity (in the case of voiced frame with pitch $p(t)$) by means of a sigmoid function. The local SNR and the harmonicity can be estimated by means of a noise estimate $M_{\hat{N}}(f,t)$ and a correlogram $A_y(f,t,p(t))$ as follows:

$$SNR(f,t) = 20log_{10}\frac{M_Y(f,t) - M_{\hat{N}}(f,t)}{M_{\hat{N}}(f,t)} \quad (1.1)$$

$$H(f,t) = A_y(f,t,p(t))/A_y(f,t,0) \quad (1.2)$$

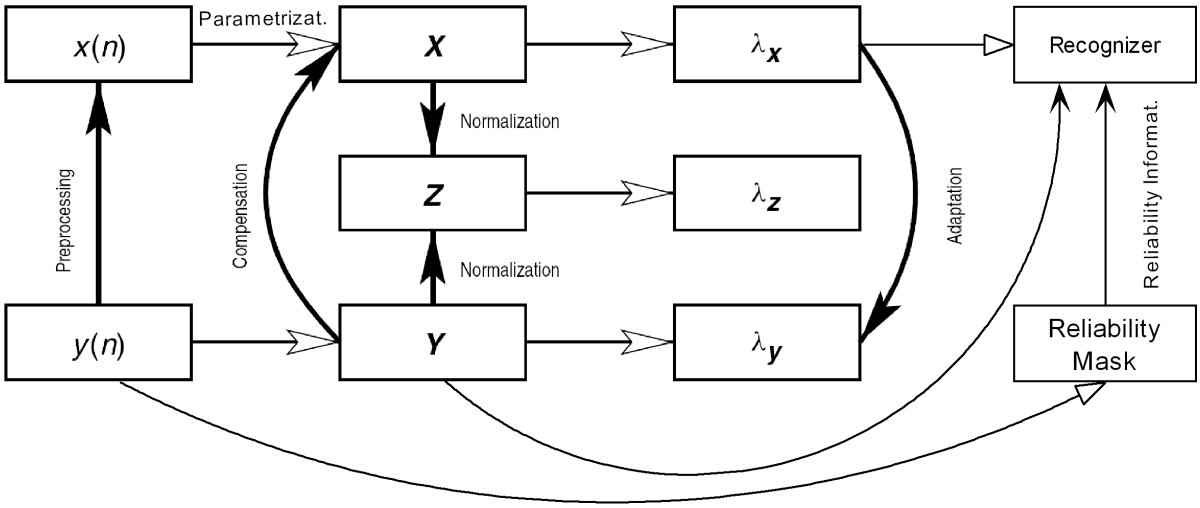


Figura 1.1: ([38] adapted) A possible classification of different conventional robust ASR techniques.

1.3. Conventional and pitch-based robust techniques

1.3.1. Conventional robust techniques

Conventional robust ASR techniques can be outlined with the scheme of Fig. 1.1 as follows:

Preprocessing: the noisy signal is cleaned or modified in temporal domain. We can mention offset and pre-emphasis in the ETSI front end [47], windows such as Hamming, SWP [29] and the variants of enhanced Wiener filter (such as in AFE [45]).

Parametrization: when a suitable acoustic representation is chosen that is robust to the speech and noise variabilities.

Compensation: the noisy features are modified to obtain an estimate of clean ones. We can mention MMSE techniques such as SPLICE [14] and VQ-MMSE Compensation [19], and the variants of SS (Spectral Subtraction) to avoid musical noise [16, 6, 24].

Normalization: when both clean and noisy representations are transformed so that the resulting features are less sensitive to noise. We can mention HEQ [13], CMN (Cepstral Mean Normalization) [35] and CTN [44].

Model adaptation: when clean models are modified to reduce the mismatch between training and testing conditions. We can mention PMC [18] and MLLR [26] .

Reliability processing: when the reliability of the noisy features is considered for recognition. We can mention WVA [7], Soft-Data [38], Multistream Recognition [8], MD (Missing Data) [11] and SFD [2].

When comparing these conventional techniques, the following conclusion can be made: Only MD technique (and its extension SFD) tends to imitate human hearing. MD does not need (for example, compared to SS) to estimate perfectly the clean or noise signals. It only needs to know the reliability mask, i. e. where speech dominates noise in the acoustic representation and vice versa. However, this technique has the default of transferring the problem to the mask estimator.

1.3.2. Robust pitch-based techniques

A bibliographic study of the pitch-based robust techniques, leads us to make the next classification:

Exploitation of harmonic structure based techniques: They do not use a pitch directly, but only some properties which derive from periodicity. We can especially mention HASE (High-lag Autocorrelation Spectrum Estimation) [43] which multiplies the high coefficients of the noisy OSA (One Side Autocorrelation) by a DDR (Double Dynamic Range) window to estimate the clean spectrum. The first 15 coefficients of the OSA are rejected because they are expected to be very contaminated by white-like noise (not correlated noise). It is also exploited the fact that in a voiced frame, spectral envelope information (short-term information) is preserved at high lags because of periodic repetitions. HASE is suitable for voiced sounds and silences, but it produces a loss of information for unvoiced frames. In order to avoid any possible mismatches, HASE is applied in both training and test. Some of our proposed techniques employ many of the HASE ideas. Another technique which exploits harmonic structure is HF [39].

Clean estimation techniques: They employ pitch extraction either to clean the signal (by means of some kind of comb filtering) or to estimate noise (with a tunnelling comb filtering) and compensate the noisy signal. As an example of the first case, WHNM ([42]) can be mentioned. An example of the second case is HT (Harmonic Tunnelling) [15]. This technique first finds the most energetic peaks of the spectrogram related to the pitch. Pitch extraction is carried out together with this peak search. An algorithm searches for the limits of the tunnelling regions which are expected to be dominated by the noise. Then, a noise spectrum estimate can be obtained by interpolating between these regions. This estimate is used in SS to obtain a clean spectral estimate. This technique has the drawback of not taking into account unvoiced frames. Another tunnelling comb techniques are FPM-NE [10] and the Frazier technique [17] which employ filters with

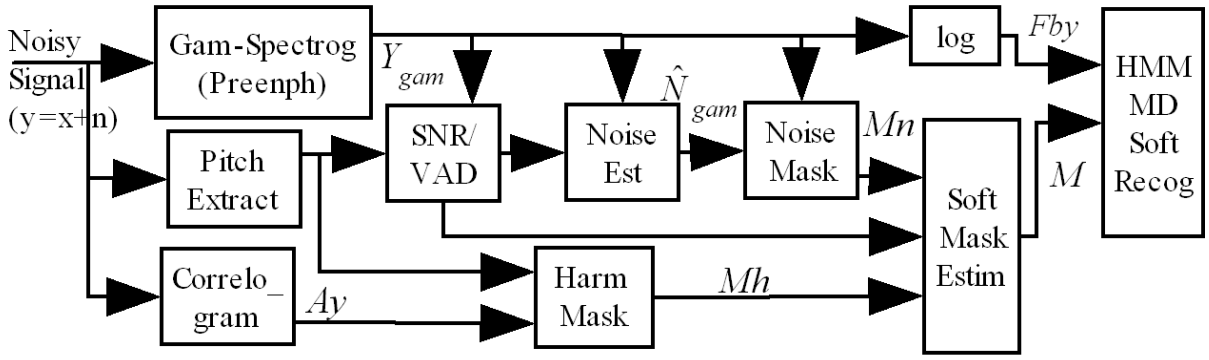


Figura 1.2: Adapted recognition system of Barker technique [3] to compare with one of our proposed techniques. Two masks are estimated, M_n based on VAD noise estimation and M_h based on the harmonicity of the correlogram. The final mask M is a combination of both masks.

impulse responses of the type of $h_T(t) = \delta(t) - \delta(t - T)$. Two of our proposed techniques are based on variants of these kind of comb filters.

Mask estimation techniques: They also employ pitch extraction to obtain a reliability mask for the considered time-frequency representation (spectrogram or cochleagram). We can especially mention the technique due to Barker [5, 3]. This technique estimates two masks, a noise soft mask M_n based on the local SNR for every time-frequency pixel estimated by means of a ten-first-frame noise estimate (Sec. 1.2), and a harmonicity soft mask M_h (based on the harmonicity of each pixel estimated by means of the noisy correlogram and the pitch, Sec. 1.2). The final mask is a linear combination of both masks. Fig. 1.2 depicts an adaptation of the Barker technique which will be compared with one of our proposed techniques. Other mask techniques have been proposed by Brown [9] and Ma [28]. This last one is based on SFD (Speech Fragment Decoding [2]) to extract the pitch and the mask of a target speaker when the noise is another speaker.

Doing a fair comparison of above pitch-based techniques is a difficult task as we commented in the introduction (Sec. 1.1.1). Sec. 1.5 is devoted to do it. In addition to these difficulties, pitch-based techniques have others lacks:

- They do not deal with all kind of noises. For example, HASE fails with harmonic noises.
- They do not take into account unvoiced frames. For example, HT may take unvoiced frames as noise.

- They need a fine pitch estimate. For example in the case of comb filtering techniques to estimate clean signal, the spectral harmonics are not exactly located at pitch positions because of quasi-periodicity. Tunnelling comb filtering techniques to estimate the noise do not have this problem because there is «more-space» around tunnelling regions.
- In the case of proposing a pitch extractor, they involve an inaccurate pitch estimate. For example, this is the case of HT.
- They can be complex and not biomimetic. It can be observed that the more biomimetic a technique is the more efficient it is. Ma technique inspired on ASA (Auditory Scene Analysis) does not have this problem but the FPM-SE [10] does.

1.4. Proposed techniques

1.4.1. Asymmetric windows

Introduction

The asymmetric windows technique is explained in detail in a paper accepted with minor changes [34]. This technique tries to do robust ASR with low computational cost. It is inspired by the HASE technique [43] (Sec. 1.3.2), which can be interpreted as an asymmetric weighting (or windowing) of the autocorrelation coefficients of the OSA (One Side Autocorrelation). The windowed OSA is employed to obtain a clean spectral estimate and its AMFCC (Autocorrelation Mel-Frequency-Cepstral-Coeficients). Another related techniques are Cyclic-Spectrum [36], OSALPC [21], SMC [30] and LSMYWE [31] which are based on employing high-lag autocorrelation coefficients to estimate the spectrum since these coefficients are usually less contaminated by noise (Sec. 1.3.2). Another related technique which also employs asymmetric windows is that of [40], although these windows are applied in the time domain. We will only compare our asymmetric windows with HASE because HASE surpasses the other related techniques.

Recognition system

Fig. 1.3 shows the proposed ASR system to evaluate our asymmetric windows. Its front end uses very similar parameters to the ETSI FE [47]: 23 Log-Mel channels, 13-statics $(C_0, ..., C_12)$ + 13-velocity + 13-acceleration cepstral coefficients, etc.. It takes a noisy

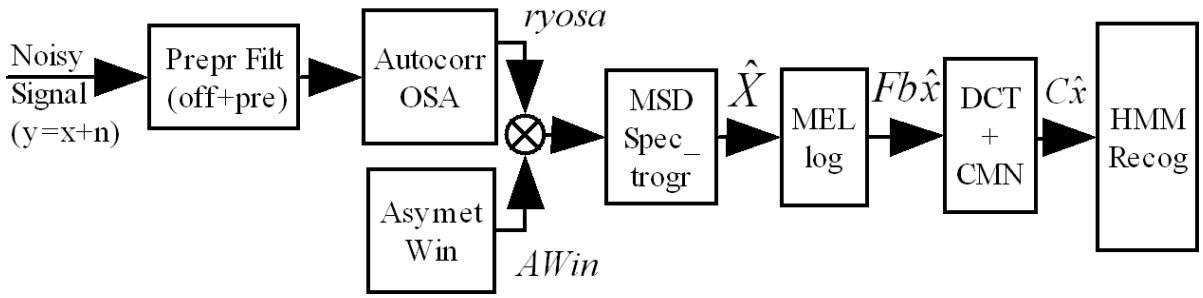


Figura 1.3: ASR system based on OSA autocorrelation with the asymmetric windows.

signal y, filters offset and enhances high frequencies, obtains the OSA of every frame and multiplies it by an asymmetric window, obtains a clean estimate of MSD (Magnitude Spectral Density) $\hat{X}$, the Log-Mel spectrum $Fb_{\hat{x}}$ and the AMFCC ($C_{\hat{x}}$). CMN (Cepstral Mean Normalization) is applied to each AMFCC and the resulting AMFCC vector is submitted to an HMM (Hidden Markov Model) recognizer. The parameters of recognizer are those of the Aurora-2 framework [37] (3 Gaussians per state, etc.). The proposed asymmetric windows are applied to both training and test in order to avoid any mismatch.

Proposed asymmetric windows

The set of proposed asymmetric windows noted as $DDR_{c,w}$ depends on two parameters: c and w (center and width in number of samples). This set is:

$$DDR_{c,w}(k) = \begin{cases} DDR_w(\frac{w}{2} - (c+1) + k) & c - \frac{w}{2} < k \leq c + \frac{w}{2} \\ 0 & otherwise \end{cases}$$
$$(k = \{0, ..., L-1\}) \tag{1.3}$$

where DDR_w is a Double Dynamic Range Hamming window [43] and L is the total window length (in number of samples) (which corresponds to OSA length). Fig. 1.4 shows an example of a $DDR_{50,250}$ applied to the OSA of a voiced frame with pitch 50 samples.

An interesting feature of the proposed windows is that they allow a variable contribution of the first autocorrelation coefficients (without discarding them completely as HASE does). Also it applies more weight to the most important coefficients by centering the window on them. Our hypothesis is that the most important coefficients for robust speech recognition are those around the pitch (or its multiples) lags because they are

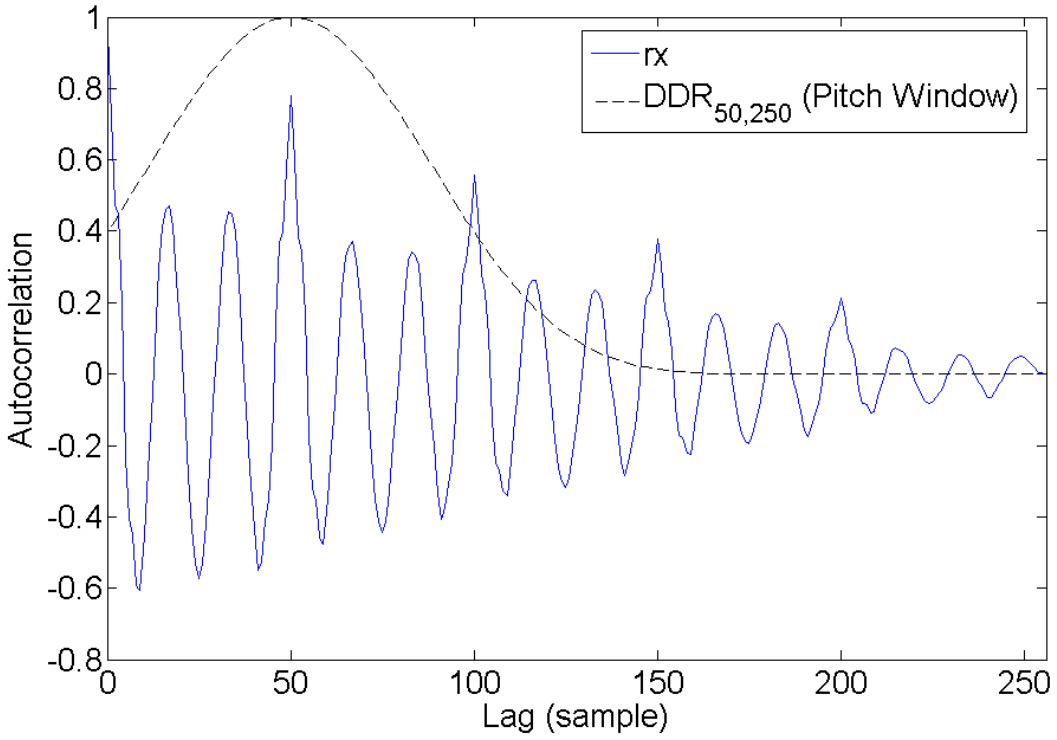


Figura 1.4: Example of a $DDR_{50,250}$ window applied to the OSA of a voiced frame with a pitch value of 50 samples.

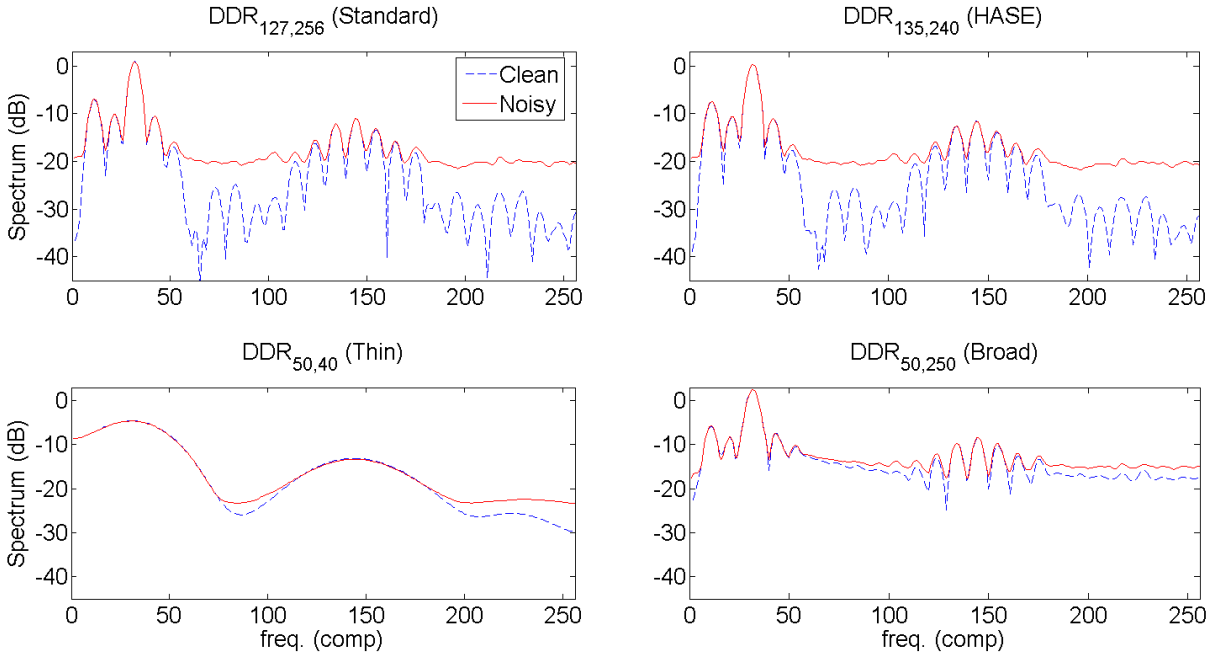


Figura 1.5: Averaged spectra of four different windows applied to a vocal with pitch=50 samples contaminated with white noise.

more energetic and less affected by the noise. In addition, they also carry spectrum envelope information. In Fig. 1.4 the asymmetric window is centered over the first pitch (lag 50). It must be taken into account that the HASE Shannon window is equivalent to our $DDR_{135,240}$.

Spectral analysis of the windows and application to unvoiced frames

Fig. 1.5 shows the clean and noisy (contaminated with white noise) spectrum of a voiced frame for four different $DDR_{c,w}$ windows. We can conclude that $DDR_{50,40}$ and $DDR_{50,250}$ have very short dynamic range (i.e. the window has not enough spectral range to cover the 80 dB necessary for speech). In spite of its short dynamic range, $DDR_{50,250}$ is quite similar to the best window for Aurora-2 that will be later obtained.

In order to avoid non homogeneous signal analysis, the same window will be applied to all types of frames (voiced, unvoiced and silence). For voiced sounds and silences, it is clear that this is always beneficial. For unvoiced it could be thought that, since lower lag coefficients (which exclusively carry the spectral envelope information) are deleted or little weighted, the use of a constant window could be harmful.

The experimental results will show that the above mentioned problems do not have effect over the system performance. In order to understand this, it is important to notice that the same asymmetric window is applied in both training and testing.

Experimental results

In order to confirm the hypothesis that the most important OSA coefficients for robust speech recognition are the pitch lag (or its multiples), a gender-dependent recognition experiment has been carried out:

Taking into account that the histogram of the average pitch per sentence (in Aurora-2 Set A) shows a mean pitch of 55 samples and two different modes for male and female speakers with pitch values at 69 and 40 samples, respectively, training and test utterances of the whole Aurora-2 (Aurora-2 Set A, B, C and clean training) are separated into three groups. These groups are: *All* (without separation depending on pitch), *P. Male* (with pitch greater than 55 samples) and *P. Female* (with pitch lower than 55 samples). A search (applying the same window in both, training and testing) for the the best window of each group is carried out by changing c and w. The WAcc (Word Accuracy in %) average (0-20 dB) results are depicted in Fig. 1.6.

It can be observed that the best windows for *All*, *P. Male* and *P. Female* groups are $DDR_{55,200}$ with 77.47 %, $DDR_{69,250}$ with 80.43 % and $DDR_{40,150}$ with 78.47 % respectively. From these results the following conclusions can be extracted:

1. For the whole Aurora-2 our proposed $DDR_{55,200}$ window with 77.47 % gives better results than the HASE window ($DDR_{135,240}$) with only 72.43 %.

2. The optimum window centers of each group just coincide with the mean pitch of each group: 55, 62 and 40 (are indicated with dashes vertical lines in the figure). This confirms our hypothesis that the most important coefficients are those around the pitch (or its multiple) values.

Tab. 1.1 shows the results obtained by the different windows tested for Aurora-2 (Set A, B and C) for different SNR values. Sec. **??** explains how the confidence intervals of the mean results are obtained. These intervals show that our results are reliable and will be only shown here and in the next table in order to avoid overloading the rest of the tables. It can be concluded that $DDR_{55,200}$ obtains better results than Hamming (very similar to ETSI FE [47]) and HASE. It can also be concluded that both the short dynamic range

1. SUMMARY OF THE THESIS: PITCH-BASED TECHNIQUES FOR ROBUST SPEECH RECOGNITION

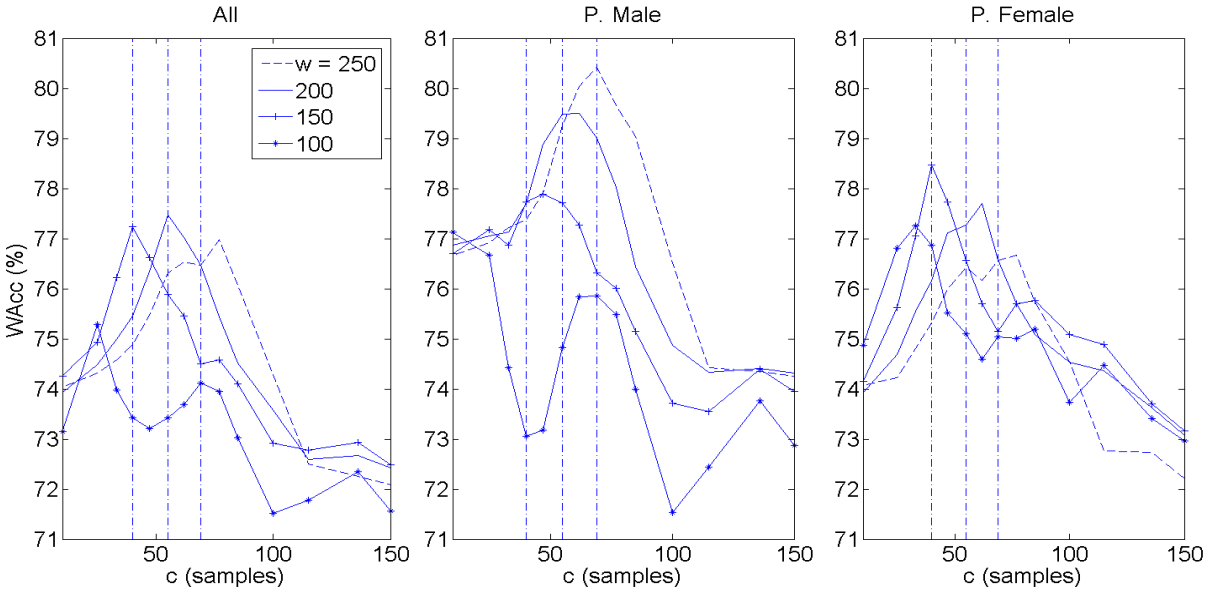


Figura 1.6: WAcc (%) for the whole Aurora-2 (0-20 dB) when all, male pitch and female pitch utterances are employed in training-test stages, againts c (center) and w (width of window). The three vertical lines correspond to the female, mean and male pitches (40, 55 and 69 samples).

Window	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Mean (20-0 dB)
Hamming (FE)	99.14	97.21	92.57	76.72	44.28	22.99	13.00	$66{,}76 \pm 0{,}80$
$DDR_{135,240}$ (HASE)	**99.15**	**97.47**	**94.37**	84.26	58.35	27.69	14.72	$72{,}43 \pm 0{,}76$
$DDR_{55,200}$ (Mean Pitch)	98.85	96.12	93.21	**85.91**	**70.00**	**42.09**	**18.07**	$\mathbf{77{,}47 \pm 0{,}71}$

Tabla 1.1: WAcc (Word Accuracies %) results obtained by different windows tested with Aurora-2 (Set A, B and C) for diferent SNR values.

Window	WM	MM	HM	Mean
Hamming (FE)	89.08	82.15	64.51	$78{,}58 \pm 0{,}64$
$DDR_{135,240}$ (HASE)	89.76	**83.16**	76.39	$83{,}10 \pm 0{,}58$
$DDR_{55,200}$ (Mean pitch)	**89.85**	82.87	**80.15**	$\mathbf{84{,}29 \pm 0{,}57}$

Tabla 1.2: WAcc results obtained by the different windows applied to Aurora-3 Spanish (real noise). WM, MM and HM mean well, medium and high mismatch, respectively.

of the proposed windows and its application to unvoiced frames are not very harmful in clean conditions as results show.

Tab. 1.2 shows the results obtained by the different windows applied to Aurora-3 Spanish (real noise) [1]. WM, MM and HM mean well, medium and high mismatch, respectively. It can be concluded that the proposed window surpasses HASE results mainly at high mismatch which is the worst condition.

1.4.2. Averaging and sifting autocorrelation

Introduction

Averaging and sifting autocorrelation estimators are explained in detail in [33]. These techniques try to estimate the clean autocorrelation of every frame by employing its pitch value. The resulting estimates are employed to obtain AMFCC features.

The averaging estimator is very related to techniques which can be reduced to a comb filter (i. e. sampling noisy spectrum at pitch harmonics). These kind of techniques are those of Kuroiwa [25], WHNM [42], etc. It is also very related to HASE [43] in the sense of supposing that the noise usually is concentrated in the first autocorrelations coefficients. We will compare our proposals with HASE.

Recognition system

Fig. 1.7 shows the proposed ASR system to evaluate different AMFCC techniques. It is very similar to that employed to evaluate asymmetric windows 1.4.1. A pitch extractor is needed to estimate the clean autocorrelation and instead of windowing the OSA, the whole (negative and positive side) the autocorrelation is employed to obtain the MSD. The window applied to this autocorrelation will be the DDR.

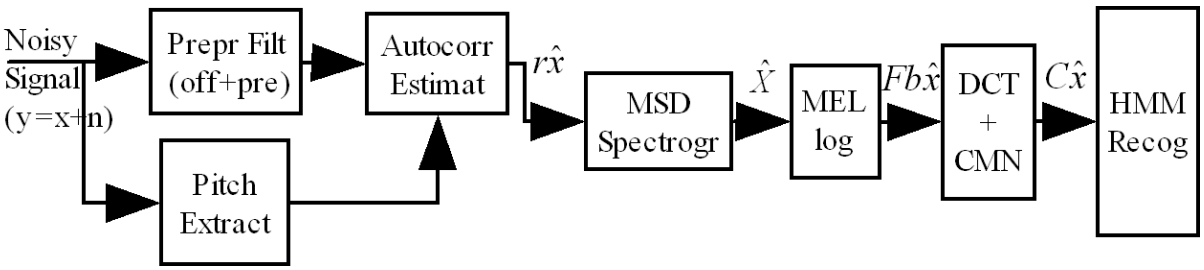


Figura 1.7: Recognition system based on the use of pitch-based clean autocorrelation estimates.

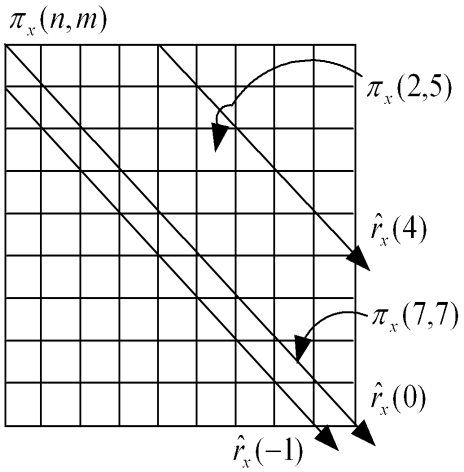


Figura 1.8: Product table for a frame $x(n)$ with 9 samples. Some products are illustrated and the diagonal arrows indicate the elements which have to be summed in order to obtain the different autocorrelation coefficients.

The pitch extractor employed here and in the following will be that presented in [33]. This pitch extractor takes the pitch provided by the ETSI xFE pitch extractor [46] and applies a smoothing processing. This smoothing is needed because the pitch provided by xFE has many errors at lows SNRs.

Product table and biased autocorrelation

The biased autocorrelation of a segment $x(n)$ is defined as,

$$\hat{r}_x(k) = \frac{1}{N}\sum_{n=k}^{N-1} x(n)x(n-k) \quad (0 \leq k < N) \tag{1.4}$$

It can be reformulated by means of a «product table» $\pi_x(n,m) = x(n)x(m)$, $(n, m =$

$0, \ldots, N-1$) (Ec. 1.5).

$$\hat{r}_x(k) = \frac{1}{N} \sum_{n=k}^{N-1} \pi_x(n, n-k) \quad (k = 0, \ldots, N-1) \tag{1.5}$$

We see that the biased coefficients can be obtained by summing diagonals of the table. Fig. 1.8 shows an example of it for a frame $x(n)$ with 9 samples. This table formulation will be useful later to better understand the proposed autocorrelation estimators.

Let's suppose now that we have a noisy signal $x(n) = p(n) + d(n)$ which is the sum of a perfect periodic clean signal $p(n)$ (which approximately represents the voiced signal) and a distortion $d(n)$ (which accounts for non-periodic components and, mainly, additive acoustic noise). If we are interested in estimating the clean periodic autocorrelation $r_p(k)$ from the noisy signal, it can be easily demonstrated that the biased estimator is not suitable because its expected value is:

$$E[\hat{r}_x(k)] = w_B^N(k) \left(r_p(k) + r_d(k)\right) \tag{1.6}$$

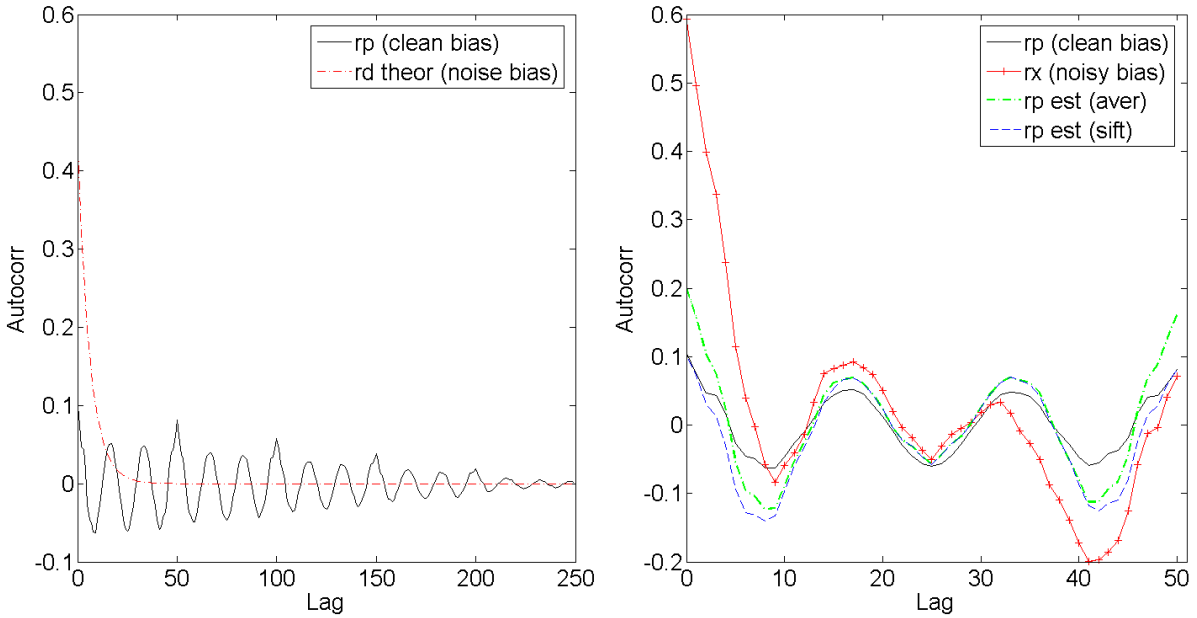
where w_B^N is a Barlett window of length N. This estimator is not robust because its error is equal to $r_d(k)$. Fig. 1.9 shows how far the noisy biased estimate is from the clean biased estimate in both, autocorrelation and spectrum domain. This illustrates the need for finding a better autocorrelation estimator.

Averaging autocorrelation

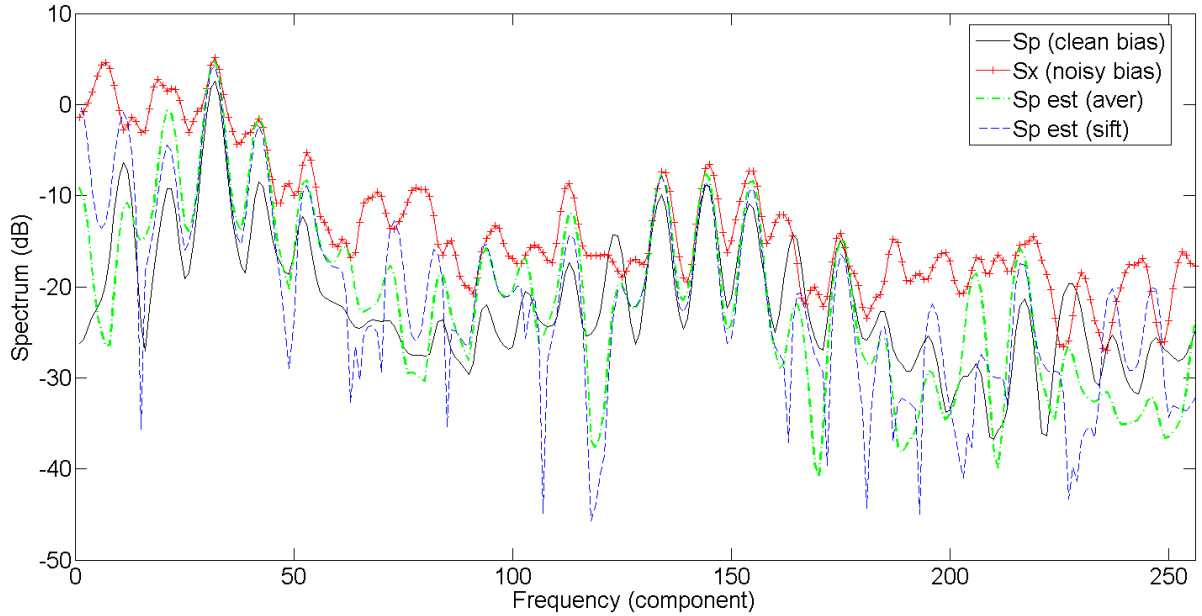
It must be noticed that if the distortion $d(n)$ was null the table would be perfect periodic and many products would be repeated. On the left of Fig. 1.10 the repeated products are marked with X for a 9-sample signal with period $T = 3$ samples. Taking this into account an estimate of the clean table can be obtained by averaging the repeated products as follows:

$$\pi_p(n, m) \approx \bar{\pi}_x(n, m) = \frac{1}{N_p^2} \sum_{i=0}^{N_p-1} \sum_{j=0}^{N_p-1} \pi_x(iT + \underline{n}, jT + \underline{m}) \tag{1.7}$$

where, for the sake of simplicity, it is supposed that there is an integer number of periods ($N = N_p * T$), $\underline{\mathrm{n}}$ is the remainder of n/T, and each averaging product $\bar{\pi}_x(n, m)$ is estimated using the idea that each clean product $\pi_p(n, m)$ is affected by a mean zero error. Fig. 1.10 shows an example of how to obtain these products. Finally, the proposed averaging



(a) Left, biased autocorrelation of the clean signal (r_p) and true AR noise autocorrelation (r_d theor) employed to contaminate it. Right, clean biased, noisy biased, averaging and sifting ($\delta = 16$) autocorrelations.



(b) Spectrums derived from clean, averaging and sifting autocorrelations.

Figura 1.9: Top, Comparison of the proposed autocorrelations for a vowel with $pitch = 50$ samples contaminated with an AR noise. Bottom, the corresponding spectra.

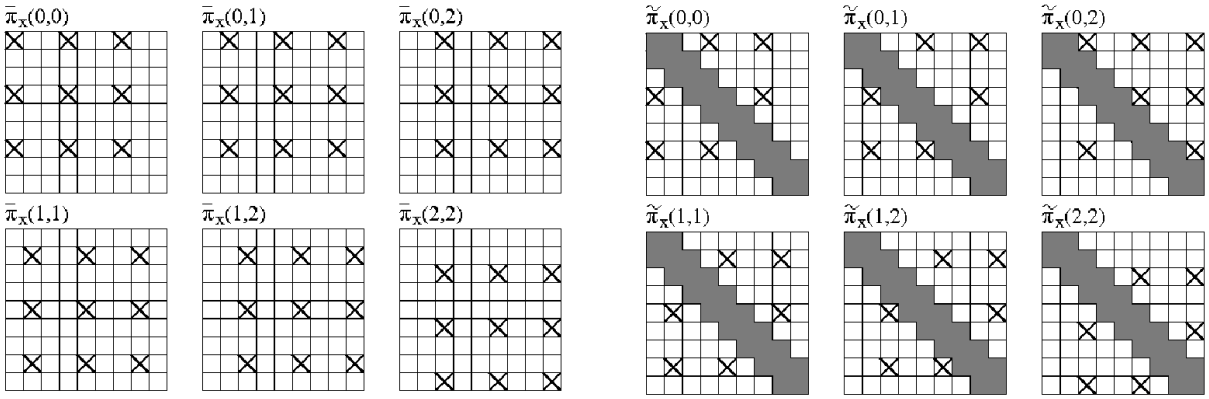


Figura 1.10: Product tables $\pi_x(n,m)$ (12 times repeated) of a $x(n)$ signal with $N = 9$ and period $T = 3$ samples. Left, computation of the different products $\bar{\pi}_x(n,m)$ for the averaging autocorrelation. Right, computation of the different products $\tilde{\pi}_x(n,m)$ for the sifting autoc. with $\delta = 2$.

autocorrelation estimator of the periodic clean signal is:

$$r_p(k) \approx \bar{r}_x(k) = \frac{1}{N} \sum_{n=k}^{N-1} \bar{\pi}_x(n, n-k) \tag{1.8}$$

It can be demonstrated (spanish Sec. **??**) that its expected value is:

$$E[\bar{r}_x(k)] = w_B^N(k) \left(r_p(k) + \frac{N_1(k)\bar{s}_d(\underline{k}) + N_2(k)\bar{s}_d(\underline{k} - T)}{N - k} \right) \tag{1.9}$$

where $\bar{s}_d(\underline{k})$ depends on $r_d(k)$ [33]. This estimator is better than the biased one because the additive error term is lower than the whole autocorrelation distortion $r_d(k)$. In particular, it can be shown that the SNR can be increased up to a factor equal to the number of available periods N_p. Fig. 1.9 shows that this estimate is closer to the clean biased autocorrelation than the biased estimate from noisy signal.

One important issue of the averaging estimation is that it can also be shown that it is equivalent to a sort of comb filtering (spanish Sec. **??**). Then, this estimator has the advantage (with respect to the biased one) of removing the noise between the gaps or tunnels placed at the middle regions of the pitch spectrum harmonics, although it does not remove noise placed at harmonics.

1. SUMMARY OF THE THESIS: PITCH-BASED TECHNIQUES FOR ROBUST SPEECH RECOGNITION

Sifting autocorrelation

Averaging estimation can be improved taking into account the HASE idea that white-like noise mainly affects to the lower lag autocorrelation coefficients. The corresponding products of these coefficients (a δ interval around the main diagonal) can be rejected or sifted to obtain a better estimate of the clean table as follows:

$$\pi_p(n,m) \approx \tilde{\pi}_x(n,m) = \frac{1}{N_\delta(\underline{n},\underline{m})} \sum_{(i,j)\in S_\delta(\underline{n},\underline{m})} \pi_x(iT+\underline{n}, jT+\underline{m}) \tag{1.10}$$

where δ is the so-called «sifting interval» and $N_\delta(\underline{n},\underline{m})$ is the number of pairs $i,j = 0,\ldots,N_p-1$ which belong to the set $S_\delta(\underline{n},\underline{m})$ (which contains the surviving index pairs). Fig. 1.10 shows how to obtain the different sifting products $\tilde{\pi}_x(n,m)$ for a $\delta = 2$.

The proposed sifting autocorrelation estimate can be obtained as:

$$r_p(k) \approx \tilde{r}_x(k) = \frac{1}{N}\sum_{n=k}^{N-1} \tilde{\pi}_x(n,n-k) \quad (k=0,\ldots,N-1) \tag{1.11}$$

It can be shown that its expected value is that of Ec. 1.9 but replacing $\bar{s}_d(\underline{k})$ by its sifted version $\tilde{s}_d(\underline{k})$ (see [33]). It can also be shown that if the noise autocorrelation is fully contained inside the sifting interval, then this estimation gives exactly the biased autocorrelation of the periodic clean signal $\hat{r}_p(k)$. Also it can be seen that sifting is the same as averaging in the interval $\delta \le k \le T-\delta$ and that sifting removes more noise than averaging in the $0 \le k < \delta$ and $T-\delta \le k < T$ intervals [33]. These intervals are just representative of the important information for ASR, i. e. the spectral envelope. Also, it can be easily seen that sifting with $\delta = 0$ becames the averaging estimator. Fig. 1.9 shows how sifting is closer to clean than averaging and that they coincide in the $\delta \le k \le T-\delta$ interval.

The important thing about the proposed estimator is that it has the advantages of the averaging (removing noise between the tunnels) plus those of the HASE technique (removing white-like noises).

Extension of sifting to silence and unvoiced frames

Sifting has been developed to estimate the clean speech autocorrelation on voiced frames. In order to avoid the use of a VAD (Voice Activity Detector) and a different estimator in silence and unvoiced frames, it will be supposed that they have a fictitious pitch of 55

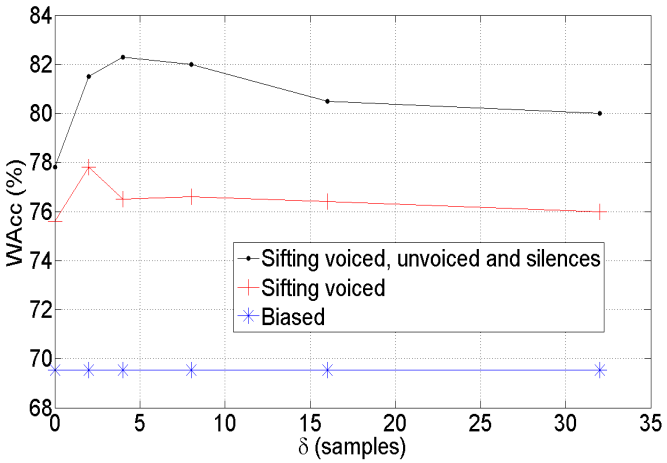


Figura 1.11: WAcc of Set-A versus the sifting interval δ when the biased autocorrelation is used for all frames ($*$), when sifting is only applied to voiced ($+$) and when sifting autocorrelation is applied to all frames $\bullet$ (voiced, unvoiced and silence).

samples which corresponds to the average human pitch (preliminary experiments showed that this is not a critical parameter of the system). In silence frames, the application of sifting is clearly suitable, but for unvoiced frames we could reasonably argue that it is not helpful but even harmful.

However, and due to similar reasons as those employed for asymmetric windows 1.4.1 the experimental results will show that this approach (the extension of sifting to all types of frames) is suitable.

Experimental results I: suitable sifting interval

Now, we will search for a suitable δ interval. Fig. 1.11 shows the WAcc (20-0 dB) results obtained for Aurora-2 Set-A versus the sifting interval for three cases: biased autocorrelation applied to all frames, sifting applied only to voiced frames and sifting applied to all (voiced, unvoiced and silence) frames. The following conclusions can be drawn:

- The sifting estimator obtains better results than the biased and the averaging ($\delta = 0$) estimators.
- It is better to apply sifting to all kind of frames than only to voiced frames. This justifies the extension of sifting to silence and unvoiced frames.
- The optimum δ is 4 samples. This value is both, large enough to reject enough contaminated products and small enough to avoid rejecting much speech information.

Technique	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Mean (20-0 dB)
A. Bias (FE)	99.06	97.65	94.74	84.06	55.30	26.53	13.63	71.65
HASE ($\delta = 15$)	99.15	97.47	94.37	84.26	58.35	27.69	14.72	72.43
A. Aver ($\delta = 0$)	99.36	97.99	95.85	89.98	72.36	36.55	12.94	78.55
A. Sift ($\delta = 8$)	98.63	96.69	94.50	89.39	76.30	44.60	14.75	80.30
A. Sift Ideal ($\delta = 8$)	98.63	97.06	95.48	91.84	82.52	61.00	29.93	85.58
AFE	99.11	97.72	96.05	91.84	82.19	59.91	28.87	85.54

Tabla 1.3: WAcc results obtained by the different techniques tested with Aurora-2 (Set A, B and C) for different SNR values.

Technique	WM	MM	HM	Mean
A. Bias (FE)	84.03	62.15	37.85	61.34
HASE ($\delta = 15$)	85.91	64.69	43.34	64.65
A. Sift ($\delta = 8$)	76.80	50.14	39.11	55.35
A. Sift Ideal ($\delta = 8$)	84.52	71.47	61.44	72.48

Tabla 1.4: WAcc results obtained by different techniques tested with Aurora-3 Danish (real noise).

In what follows, $\delta = 8$ (optimum value for the whole Aurora-2 [33]) will be taken as our optimum sifting interval.

Experimental results II: Aurora 2 and 3

Tab. 1.3 shows the results for the different autocorrelation estimators, HASE and the ETSI AFE front-end [45] over Aurora-2. It can be observed that the application of sifting to unvoiced frames is not very harmful as clean results show. In general, sifting surpasses all except AFE results because this is a more sophisticated front-end which brings together different robust techniques. Sifting with ideal pitch (i. e. pitch extracted from the corresponding clean signal) could perform as well as AFE as shows in the *A. Sift Ideal* row.

Tab. 1.4 shows the results obtained over the real noise database Aurora-3 (Danish). It can be observed that sifting would require a better pitch extractor to improve the HASE results. In this case, sifting could surpass HASE in more than 18 % of WAcc (*A. Sift Ideal* experiment).

Technique	Set A				Set B				Set C		Mean (20-0 dB)
	Subw	Babb	Car	Exhi	Rest	Stre	Airp	Trai	Subw MIRS	Stre MIRS	
A. Aver ($\delta = 0$)	79.19	80.14	77.36	76.54	81.03	79.08	80.73	78.73	75.63	77.01	78.55
A. Sift ($\delta = 8$)	83.62	81.96	80.56	80.80	78.45	82.15	80.16	80.63	76.16	78.47	80.30
A. Sift ($\delta = Ideal$)	89.07	87.49	86.68	86.88	85.03	88.07	85.92	86.03	85.17	85.96	86.63

Tabla 1.5: WAcc results obtained by the different techniques tested with Aurora-2 (Set A, B and C) for different SNR values.

Experimental results III: dynamic sifting

Tab. 1.5 shows the WAcc over Aurora-2 depending on the type of noise. It is observed that sifting surpasses averaging for all noises except for *Restaurant* and *Airport*. There are several reasons for this shortcoming such as errors in pitch extraction or a unsuitable δ. Another experimental results have shown that with other δ values (not 8), this shortcoming with *Restaurant* and *Airport* can be sorted out.

This points out the need of applying sifting with a dynamic value for δ(that is, a suitable value for each instant or utterance). *A. Sift* ($\delta = Ideal$) is an oracle experiment which selects the best δ for each utterance. It shows the limits of improving the results by means of a dynamic delta for each utterance. Thus, dynamic sifting is a possible future reasearch line.

1.4.3. Pitch-based noise estimation

Introduction

Our proposed pitch-based noise estimation technique is explained in detail in [32]. Noise estimation is an important issue in robust speech recognition and there exit many approaches to do it. If you want to perform this task, taking into account the spectral masking effect [48], the only way to do it is by interpolating noise from regions where it is known. VAD noise estimators [38] do this and are suitable for stationary noises. Other techniques, such as those which can be reduced to a comb filtering of noise, can be employed in order to obtain more regions of noise and to face non-stationary noises. **HT** (Harmonic Tunnelling) [15] is an example of these kind of comb techniques which require

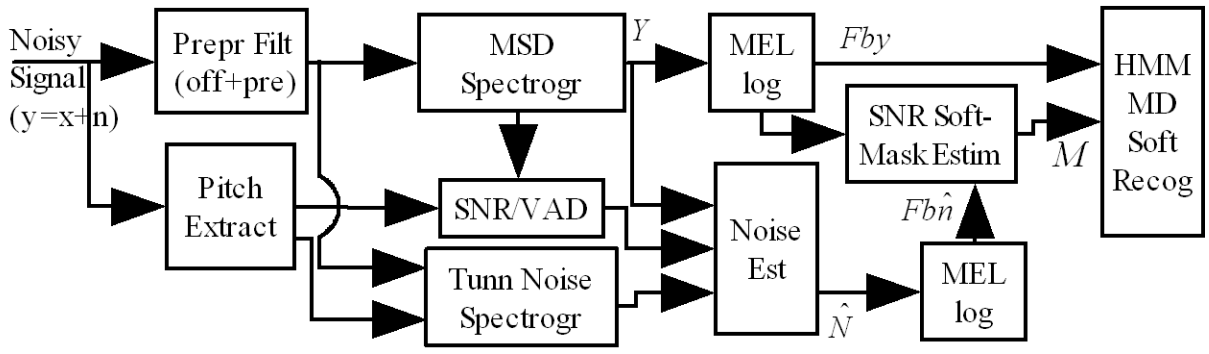


Figura 1.12: Proposed recognition system to evaluate MD ASR from pitch-based noise estimation.

.

a pitch extractor. Here we propose a noise estimate which combines VAD estimates and a modification of HT noise estimates by means of the pitch extraction. In addition to the modifications applied to HT (such as avoiding overestimation and not including unvoiced frames as noise) the important contribution of our proposal is that it fully exploits pitch information to perform robust ASR as we will see in Sec. 1.5.

The proposed noise estimate will be evaluated on SS (Spectral Subtraction) and MD (Missing Data) [11]. It will be also compared with a VAD noise estimate and with an adaptation of the Barker's technique [3] which also employs MD and pitch.

Recognition system

Fig. 1.12 shows the proposed MD system to evaluate the proposed noise estimation in ASR. It is very similar to that employed for sifting 1.4.2.

The *SNR* (global Signal to Noise Ratio estimator of the utterance) and *VAD* block take as inputs the noisy MSD (Magnitude Spectral Density) Y and the pitch. The *Tunnelling Noise Spectrogram* block estimates the noise in voiced frames using a modification of the HT technique which makes use the of noisy signal and the pitch estimates. Our center block *Noise Estimator* takes Y, *SNR*, *VAD* and the tunnelling noise estimate to provide a spectrogram noise estimation $\hat{N}$. Y and $\hat{N}$ are the inputs to the MEL filter bank and the log compressor (which yields Fb_y and $Fb_{\hat{n}}$). These two last outputs are used to estimate an SNR of every frequency-time pixel and then the corresponding soft mask M. Finally, M and Fb_y are employed by the *MD Soft Recognizer* [4]. The parameters of the recognizer are those commonly employed over Aurora-2 for ASR with spectral features (9 Gaussians per state, [3]).

Now we will describe the most important blocks of the proposed system. Note that the different parameters were determined through preliminary experiments performed over a set of training (not testing) sentences of Aurora-2 contaminated with noise.

VAD based on pitch

The proposed VAD is based on the «main source model» of speech (Sec. 1.2) because once the pitch (main source) is located, the remaining speech sounds can be localized too.

Our VAD detects three different classes of frames: voiced, unvoiced and silences. Frames labeled as voiced correspond to frames where the pitch extractor gives a valid pitch. Unvoiced frames are searched in an interval of 20 frames before or after a sequence of voiced frames and identified when the instantaneous SNR of high frequencies is greater than 3 dB:

$$\hat{SNR}^{HF}(t_k) = 10 * log_{10}(E^{HF}_{\hat{X}}(t_k)/E^{HF}_{\hat{N}}(t_k))E_{\hat{N}}(t_k)) \tag{1.12}$$

$$\text{where } E_S(t_k) = \sum_{j=j_{1,8KHz}}^{j_{4KHz}} |S(\omega_j, t_k)|^2 \tag{1.13}$$

The reasons for this condition is that unvoiced sounds never occur in isolation and their energies are mainly between 1800 and 4000 Hz (sample frequency) [41]. The clean spectrogram $\hat{X}$ is estimated through the noise estimate $\hat{N}$ based on the 10 first-last noisy frames. Subsequent experiments have also shown that at low SNRs, this unvoiced estimation takes many noise frames as unvoiced. So when the estimate of the global SNR is less than 10dB, it is assumed that unvoiced signals are mixed with noise and no detection of unvoiced frames is carried out. This global SNR is estimated by means of $\hat{X}$ and $\hat{N}$.

Silence frames are those which have been classified neither as voiced nor unvoiced.

VAD Noise Estimate

NVAD (VAD noise) is estimated by interpolating the noise from silence (noisy) frames. An averaging of the noisy MSD Y of the closest 10 silence frames gives the estimate in each voiced or unvoiced frame.

Harmonic Tunnelling Noise Estimate

The continuous MSD of a noisy signal $y(n)$ with N samples at frequency ω is:

$$Y(\omega) = |\frac{\sum_{n=0}^{N-1} y(n)w(n)e^{-i\omega n}}{\sqrt{N}}| \tag{1.14}$$

where $w(n)$ is the Hamming window. Then, the discrete **NTun** (a variation of harmonic tunnelling noise) is estimated by interpolating tunnelling samples $Y(\omega_l)$ which are obtained from the pitch frequency (ω_0) as follow:

$$\begin{aligned} \hat{N}_{tun}(\omega_j) &= Interp(\boldsymbol{\omega_l}, Y(\boldsymbol{\omega_l}), \omega_j) \\ \omega_l = \omega_0(l + \frac{1}{2}),\ l &= \{-1/2, 0, 1, 2, .., ceil(\pi/\omega_0)\} \\ \omega_j = \frac{2\pi j}{NFT},\ j &= \{0, .., NFT/2 - 1\} \end{aligned} \tag{1.15}$$

Figure 1.13 shows an example of tunnelling noise estimation. *NTun* has the problem of overestimation mainly at high SNRs (more than 10dB) because of the spectral window (as shown in the figure at low/high frequencies).

VAD+Tun Noise Estimate

The final noise estimate is *NVAD* but corrected, depending on global SNR estimate, at voiced frames as follows:

- If global $SNR < 10dB$: NVAD is replaced by *NTun*.
- Otherwise: *NTun* is used as an upper bound for *NVAD*.

The reason for using *NTun* only as an upper bound when $SNR \geq 10dB$ is that overestimation is more likely in this case. Also, real noises tend to be more stationary at high SNRs [27]. The final noise spectrogram *NVADTun* is smoothed and its $Fb\hat{n}$ spectrogram (Filter bank Mel-Log representation) is obtained. Fig. 1.14 depicts a comparative example.

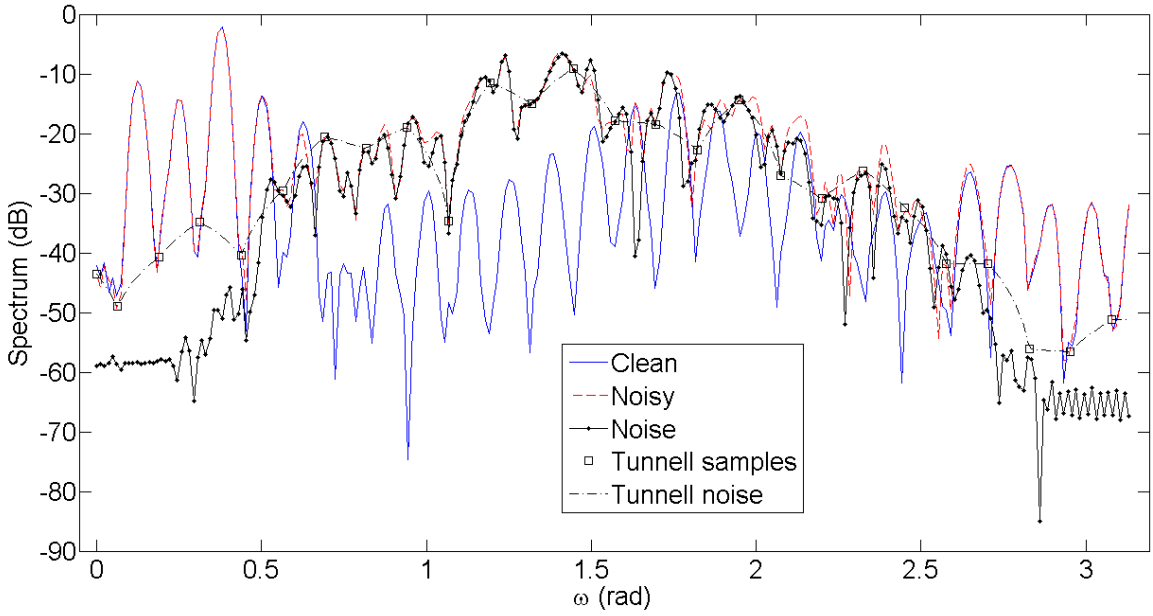


Figura 1.13: Example of tunnelling noise estimation on a voiced noisy frame with pitch $\omega_0 = 0.126$ rad..

Mask Estimation

The clean spectrogram $Fb\hat{x}$ is estimated subtracting Fby and $Fb\hat{n}$ and then the local SNR of every pixel (mel filter ch_j at time t_k) can be obtained as:

$$\hat{SNR}(ch_j, t_k) = 20 * log_{10}(e^{Fb\hat{x}(ch_j,t_k)}/e^{Fb\hat{n}(ch_j,t_k)}) \tag{1.16}$$

This is passed through a sigmoid function to obtain the soft mask estimate M (reliability values between [0, 1]). The threshold and the slope of the sigmoid are -3 dB and 0.2 respectively and they have been determined empirically.

Experimental results

Tab. 1.6 shows the WAcc results with Aurora-2. The first four systems use the cepstrograms with CMN (*Ceps*). *FE* stands for a cepstrum obtained from the spectrogram Fby and provides a very similar result to the ETSI front-end [47], *AFE* is the ETSI front-end [45], and *A. Sift* is the sifting autocorrelation (Sec. 1.4.2) which is an example of pitch-based robust technique. *N. VAD+Tun, SS* is when the proposed noise estimate is used in an Cepstral SNR-dependent SS (Spectral Subtraction) scheme which parameters have

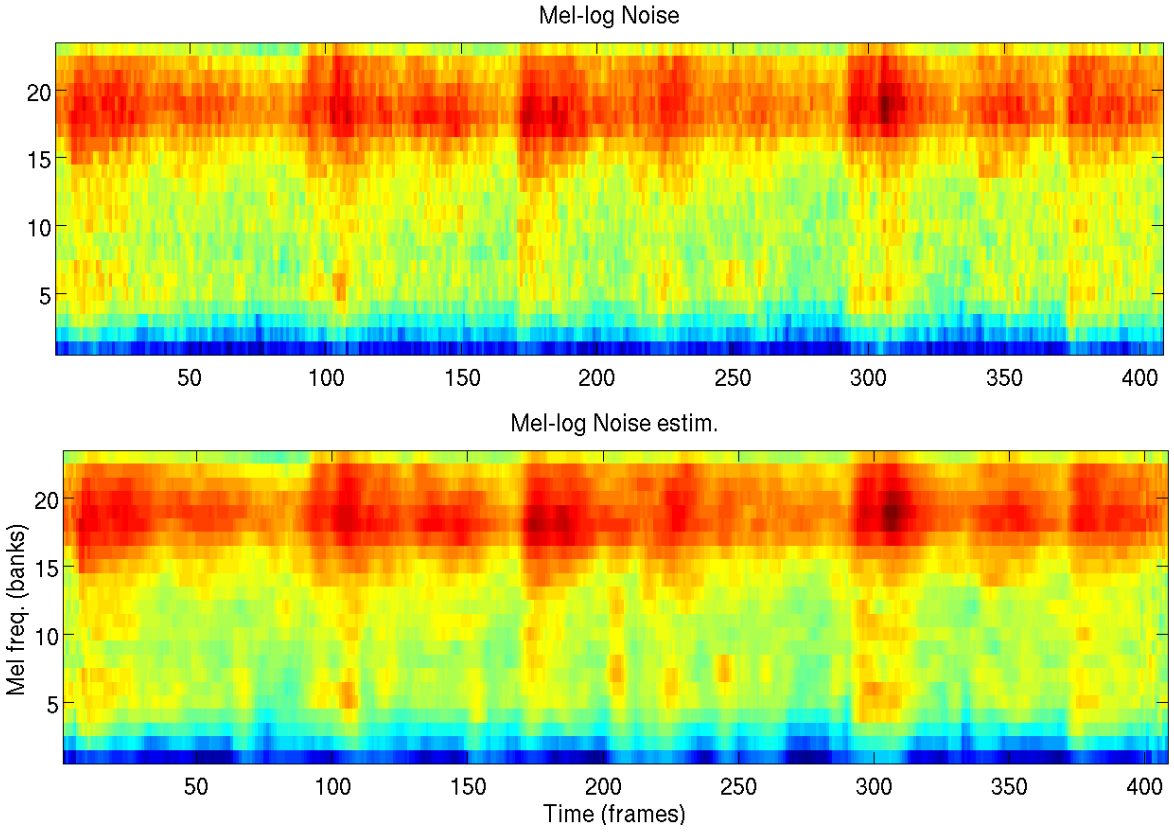


Figura 1.14: Subway Mel-log noise and its estimation from Aurora-2 utterance 4460806 at 0dB

System	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Mean (20-0 dB)
FE (Ceps)	99.14	97.21	92.57	76.72	44.28	22.99	13.00	66.76
N. VAD+Tun, SS (Ceps)	99.36	96.66	92.09	81.84	64.09	37.06	9.72	74.35
A. Sift (Ceps)	98.63	96.69	94.50	89.39	76.30	44.60	14.75	80.30
AFE (Ceps)	99.11	97.72	96.05	91.84	82.19	59.91	28.87	85.54
N. VAD+Harm (MD, Cocl)	98.67	96.18	92.67	84.17	74.21	50.41	17.65	79.53
N. VAD (MD)	98.76	96.19	93.38	88.42	77.92	49.52	15.56	81.09
N. VAD+Tun (MD)	98.78	95.79	92.04	86.66	78.03	54.43	18.40	81.39
N. VAD+Tun Ideal (MD)	98.78	95.97	92.81	88.57	84.24	74.43	55.83	87.21

Tabla 1.6: WAcc results obtained by different systems tested with Aurora-2 (Set A, B and C) for different SNR values.

been optimized to avoid musical noise.

The next four systems estimate a soft mask to recognize (*MD*). *N. VAD*, *N. VAD+Tun* and *N. VAD+Tun Ideal* use our proposed noise estimates. *Ideal* means that pitch is obtained from corresponding clean signal. These three systems employ a 23-channel spectrogram as acoustic representation. However, *N. VAD+Harm*, which is an adaptation of Barker's technique explained in Sec. 1.3.2 especially developed to compare with our technique, employs a 23-channel cochleagram (Cochl). Its VAD is the same as the one we have previously proposed but adapted to the cochleagram representation. The values of threshold and slope of the sigmoid functions of M_n and M_h are (-6 dB, 0.8) and (0.8,70) respectively, and they have been determined empirically.

The following conclusions can be drawn:

- *N. VAD+Tun* performs better in Spectral MD than in Cepstral SNR-dependent SS. This is because SS is more sensitive to errors of noise level. This is the reason why MD is preferred instead of the SS approach as HT does.

- If we compare *N. VAD* with *N. VAD+Tun*, we see that the addition of *NTun* provides benefits, mainly at low SNRs. However, we also see that tunnelling is not beneficial at higher SNRs. This can be understood if we take into account that Aurora-2 mainly consists of (quite) stationary noises. On the other hand, we think that our technique can be more helpful for non-stationary or sporadic noises.

- If we compare *N. VAD+Harm* with *A. Sift* and *N. VAD+Harm Cocl*, it seems that the proposed noise estimate makes a better use of the pitch information than the other two. However, this can not be concluded definitively as several causes can be influencing on this. Among others, that *A. Sift* and *N. VAD+Harm Cocl* can be more sensitive to pitch errors or that their parameters are not optimally tuned. This kind of problems shows the need of determining which technique makes a better use of the pitch information. The answer to this question will be addressed in Sec. 1.5.

- *N. VAD+Tun Ideal* show that with a better pitch estimation, results could be considerably improved (overcoming *AFE*). In future work (Sec. 2.3) different possibilities to improve the pitch estimation are discussed.

1.5. Equivalences and limits of the pitch-based techniques

1.5.1. Basic mechanisms and equivalences

Voiced basic mechanisms

In previous sections we have studied and proposed different pitch-based techniques for robust ASR. Now, we will compare them in a fair way by means of using some equivalences. In principle, they can be supposed as different if we only pay attention to some specific details (pitch extractor, processing of unvoiced and silence frames, etc.). However, they can be reduced to one of these four basic mechanisms which depend on the robust method applied to voiced frames:

1) **Exploitation of the harmonic structure**: these mechanisms do not require a pitch extraction but only some properties which can be derived from periodicity. SWP [29], HASE [43] and Asymmetric Windows (Sec. 1.4.1) try to «clean» the signal using these properties. HF [39] estimates the noise by exploiting the spectral harmonic shape.

2) **Comb estimation of clean signal**: these mechanisms use the pitch frame to apply some kind of comb filtering, i. e. some kind of algorithm which can be reduced to a sort of removing noise between the gaps (or tunnels) which are in the middle between the pitch spectrum harmonics. The resulting clean signal can be recognized from its cepstral representation. WHNM [42], PHCC [20] and Sifting (Sec. 1.4.2) use these mechanisms.

3) **Tunnelling estimation of noise**: these mechanisms are the opposite of the preceding ones and estimate noise (tunnelling noise) employing tunnelling samples, that is, the spectral gaps between the harmonics. The resulting noise estimate can be employed in SS, MD, etc.. HT [15], FPM-NE [10] and Pitch-based Noise Estimation (Sec. 1.4.3) use these mechanisms.

4) **Harmonicity mask estimation**: this mechanism estimates the mask of each frequency-temporal pixel by means of the correlogram and the pitch. Cochleagram techniques related with ASA, such as the adaptation of Barker's technique (Sec. 1.3.2) and the Ma's technique [28] employ this mechanism.

Taking into account these mechanisms we can investigate about which is the best one and whether they fully exploit the pitch information to improve the recognition in voiced frames. These questions are answered in Sec. 1.5.2.

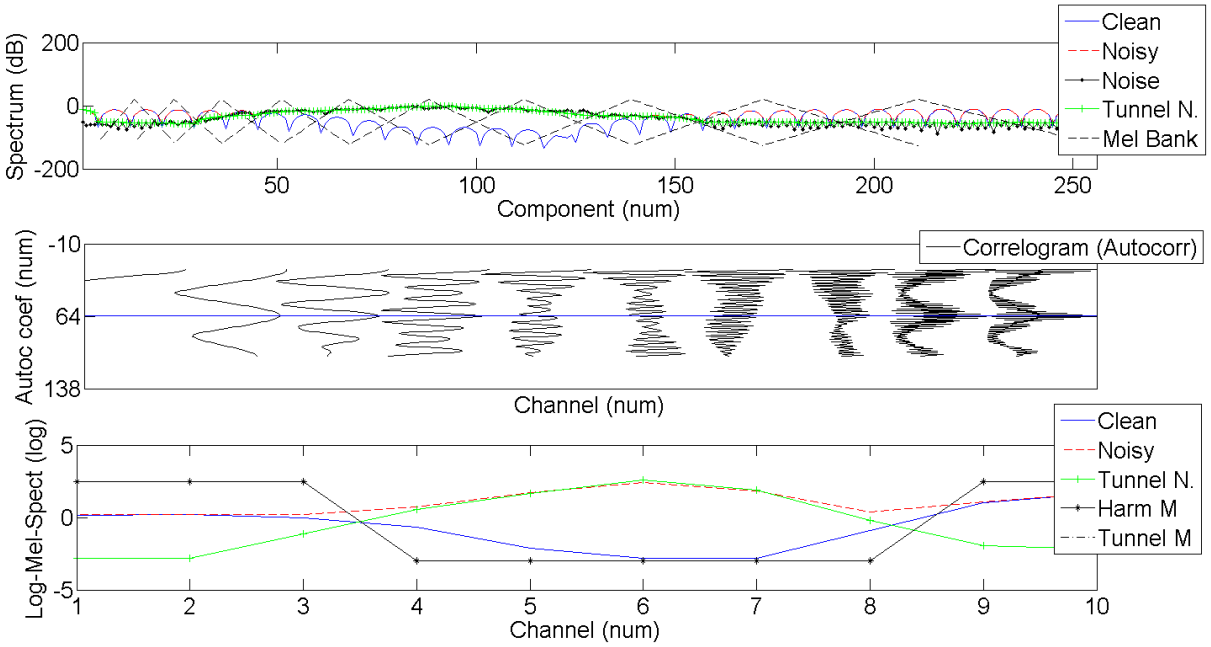


Figura 1.15: Comparison of the mechanisms to estimate a tunnelling mask and a harmonicity mask. Both masks are shown in the Log-Mel Spectrum plot

Comparing tunnelling and harmonicity masks

It can be shown that the mask derived from tunnelling noise is similar to that derived from harmonicity measures if similar channel numbers and a suitable selection of thresholds are applied.

Fig. 1.15 can help to understand this similarity. The clean and tunnelling noise estimate, which indicates where the mask should be 1 or 0, are on top of the picture along with the 10 Mel filter bank, employed in tunnelling estimation. The outputs of the 10 gammatone channels of the correlogram employed to estimate harmonicity mask are in the middle plot. The two mask estimates (*Harmonicity and Tunnelling Mask*) are overlapped at the bottom of the picture along with the Log-Mel spectra employed to estimate the tunnelling mask, showing the strong similarity of both estimates. We can conjecture that both masks will yield similar recognition results (hypothesis H1).

1.5.2. Optimum voiced mechanisms

Optimum pitch-based noise estimation

Let's suppose that we have a noisy signal $x(n)$ of length N which is the sum of a pure periodic clean signal $p(n)$ and a distortion $d(n)$. T (or ω_0 in radians) is the period of $p(n)$ and, for the sake of simplicity, we also suppose that we have an integer number of periods N_p ($N = N_p * T$). Its complex discrete noisy spectrum is:

$$X(\omega_k) = P(\omega_k) + D(\omega_k) \qquad (k = 0, ..., N-1) \tag{1.17}$$

Taking into account the periodicity of $p(n)$, the above equation can be expressed as follows:

$$X(\omega_k) = \begin{cases} P(\omega_k) + D(\omega_k) & \text{if} \quad \omega_k = \omega_0 m \\ D(\omega_k) & \text{otherwise (tunnelling samples)} \end{cases} \tag{1.18}$$

where $m = 0, 1, .., T-1$. From this equation, we can deduce that only a percentage $(Np-1)/Np$ of the N noise spectral samples can be recovered if we only know the pitch period T, no matter how the noisy signal is transformed. The remaining noise frequency samples are mixed with the speech harmonics and can not be recovered, although they can be estimated by applying some type of interpolation.

We can consider that the noise spectrum estimates obtained from tunnelling samples and interpolation are optimal in the sense that minimal assumptions about the noise are required (only an interpolation model). In practice, it must be also taken into account that the resulting noise estimation has some problems like non perfect periodicity or unavoidable time-window which also widens the harmonics. The reason of only taking one tunnelling sample (between the harmonics) in the proposed Pitch-based Noise Estimation technique is this widening.

Optimum voiced mechanisms

Let us consider the following three points:

1. Tunnelling noise estimate is theoretically optimum (just argued above).
2. The similarity between tunnelling and harmonicity masks (Sec. 1.5.1).
3. MD (with ideal mask) provides much better results than other techniques which employ a noise estimate (such as SS) (Sec. 1.3.1).

Technique	Mean (20-0 dB) [0 dB]		
	Technique «per se» (without oracle)	Oracle mask unvoc. and sil.	Oracle mask all
FE (Spectr.)	33.30 [7.66]	64.25 [25.04]	95.01 [90.18]
$DDR_{55,200}$ (Spectr.)	35.84 [5.84]	73.16 [37.98]	90.35 [82.75]
A. Sift ($\delta = 8$) (Spectr.)	36.61 [8.09]	77.92 [47.72]	93.36 [88.94]
N. VAD+Harm (Cocl.)	85.95 [72.21]	89.15 [73.13]	95.11 [89.40]
N. VAD+Tun (Spectr.)	87.21 [74.43]	90.87 [79.46]	95.01 [90.18]

Tabla 1.7: WAcc results for the whole Aurora-2 (Set A, B and C) obtained by four techniques which represent the four basic voiced mechanisms. 0 dB result is shown in bracket. Ideal pitch is employed.

From these three considerations, we can say that mask estimation mechanisms based on tunnelling or harmonicity, along with MD recognition, provide a very solid framework for pitch-based recognition of voiced frames, and that in ideal conditions these can be considered as an optimum mechanisms (hypothesis H2).

Experimental results

In order to compare the robustness of the four basic mechanisms for voiced frames, WAcc results in spectrogram (or cochleagram) domain, with ideal pitch and with oracle mask in unvoiced and silence frames for different techniques (representative of each mechanism) are shown in Tab. 1.7.

FE is used as baseline (no robust). $DDR_{55,200}$ corresponds to the asymmetric window (Sec. 1.4.1) and represents the mechanisms based on exploiting the harmonic structure. *A. Sift* corresponds to the sifting autocorrelation technique (Sec. 1.4.2) and represents the mechanisms based on comb estimation of the clean signal. *N. VAD+Harm* is the adaptation of Barker's technique (Sec. 1.3.2) and represents the mechanisms based on harmonicity mask estimation. *N. VAD+Tun* is the tunnelling mask (Sec. 1.4.3) and represents the mechanism based on tunnelling noise estimation.

The first column shows the results obtained by these techniques (all-ones mask has been employed for the first three techniques). The second column shows the same experiments but applying oracle masks to unvoiced frames and silences (this shows the success of the voiced mechanisms), and third column shows oracle mask results. The soft-mask threshold and slope of *N. VAD+Harm* and *N. VAD+Tun* have been re-optimized to improve the results in the second column.

It can be concluded that the best voiced mechanisms are the two last ones, i. e. harmonicity and tunnelling mask estimations. Their results are quite similar although tunnelling is a bit better. This increment can be due to the difference between the Mel scale of the spectrogram and the ERB scale of the cochleagram. Except for this difference, it can be said that these mechanisms are similar and that they are best ones. This confirms many of the previous statements made in this section (hypothesis H1 and H2).

1.5.3. Limits in pitch-based recognition

Performance limits

If we compare the first and second columns of Tab. 1.7 for the proposed technique *N.VAD+Tun* and it is taken into account that second column contains an approximation to the best performance that we can obtain with the pitch-based techniques (because unvoiced and silence frames have oracle mask and voiced frames have one of the optimum voiced mechanisms) we can conclude that the proposed pitch-based noise estimation technique (first column) is almost optimum because its results are not very far from this upper boundary results (second column).

Let us compare now the second and third columns of the table. Although the results of the second column are not very far from those of the third one (oracle masks for all frames), we can see that the pitch-based mask estimation methods will never perform as well as the oracle masks (this is specially clear at 0 dB), independently of the accuracy of the pitch extractor employed. This points out that in order to obtain further improvements, more information than that extracted from the pitch trajectories would be required to approximate the performance of the oracle masks. This extra information could be obtained from the noise itself or accurate speech models.

Recognition of speech without pitch

This thesis has been devoted to the recognition of speech as it is usually uttered, that is, with vibration of the vocal folds. However, speech can be sometimes emitted without pitch (whispered speech, [49]) or with multiple pitch values (vocal harmony, in music). Humans can recognize these voices even in noise conditions. This can create the illusion that pitch is not an important cue in robust speech recognition. However, as it is explained in the introduction section, although we consider the pitch as an important cue, it is not the only one. We consider the ASR of whispered speech as an important field for future work

which we are willing to study. To do that, the following ideas could be considered (most of them extracted from this Thesis):

- Design of a VAD detector similar to that developed in Sec. 1.4.3, taking into account the main source model of speech. In this Thesis, the main source is associated to pitch. Now, the main source could be localized where instantaneous SNR is higher (whispered) or multiple pitches rise at the same time (vocal harmony).
- Adaptation and improvement of the models for this type of speech, taking into account that now it has a flatter spectrum, with less energy (whispered), etc. [49, 23].
- Application and adaptation of the MD (or SFD [2]) techniques to this type of speech.

1. SUMMARY OF THE THESIS: PITCH-BASED TECHNIQUES FOR ROBUST SPEECH RECOGNITION

Capítulo 2

Conclusions, Contributions and Future Work

2.1. Conclusions

The present work is motivated by the need of proposing and carrying out a comparative study of robust speech recognition techniques based on pitch (not including robust pitch extraction). The main conclusions are summarized below:

- Taking into account that the message of a speech signal is coded by means of three kind of elements (voiced sounds, unvoiced and silences) and the way they are combined, we can say that the speech signals «mainly» consists of voiced sounds which are surrounded by the unvoiced sounds. This has been referred to as «main source model» which is a simplify definition of speech that it has been employed to develop a VAD (Sec. 1.4.3). This model is also suitable for whispering speech if a noise is taken as the main source.

- The state of the art of conventional techniques for robust ASR leads to the conclusion that MD (Missing Data) techniques can obtain very high performances (close to human) without the need of perfectly estimating the noise or the clean signal. However, this transfers the problem to the mask estimation block.

- The comparative study of the pitch-based techniques found in the bibliography (exploitation of harmonic structure, clean signal estimation and mask estimation techniques) is a difficult task because each author employs a different pitch extractor, each technique uses extra techniques and sometimes it is not clear if the author is

proposing a new pitch-based robust technique or a robust pitch extractor itself. Even so, we have tried to establish some equivalences between the different techniques and the recognition limits of the pitch-based techniques.

- A set of asymmetric windows called $DDR_{c,w}$ has been proposed which extends the HASE technique [43] that is employed to perform robust feature extraction by means of the OSA in white-like noises (contained in the first autocorrelation lags). It has been shown that the highest robustness is obtained by windows centered around the pitch values because these are the most energetic autocorrelation lags (have more SNR) and preserve the formant information. The coefficients which should be less weighed are the first ones because they are the most affected by the noise.

- A clean autocorrelation estimation method called *sifting* (based, in turn, on another proposed estimator, which was referred to as *averaging* estimator) has been proposed. It uses the pitch and depends on the sifting parameter δ which informs about the amount of autocorrelation products which are rejected because they are supposed to be more contaminated by noise. It has been shown that, taken a suitable δ value, which includes the first (more energetic) autocorrelation coefficients of a white-like noise, the estimate can be equal to the clean signal autocorrelation under certain assumptions.

- Taking into account that for $\delta = 0$ sifting is a sort of comb filtering (a spectral sampling of noisy signal at the pitch harmonics) and that many of the pitch-based techniques can be reduced to a comb filtering, we can concluded that sifting is an extension of many of these comb techniques. Sifting has the advantages of the comb techniques (eliminating the noise placed between pitch harmonics) and HASE (eliminating white-like noises).

- The extension to unvoiced frames of both the $DDR_{c,w}$ windows and sifting could degrade the performance (mainly at clean conditions) because the information of unvoiced sounds is mainly contained in the first autocorrelation coefficients, which tend to be removed. Nevertheless, this problem can be avoided by applying the same technique in both, training and test stages.

- Techniques such as HT [15] or that of Frazier [17], based on estimating the noise spectrum in voiced frames by means of tunnelling samples (spectral samples which are between the pitch harmonics), have the problem of including as noise unvoiced

frames (VAD is not used) and of overestimating it, degrading the performance as they also employ SS (Spectral Subtraction) which is very sensitive to these overestimations. In order to avoid these problems a recognition system, which includes a VAD+Tunnelling noise estimation and MD instead of SS, has been proposed.

- The proposed VAD uses the pitch location in order to locate the rest of the speech elements taking into account the *main source model* of speech. The tunnelling estimate also uses the pitch so we have finally proposed a *noise estimation based completely on pitch*.

- If we do not consider some elements of the pitch-based techniques, such as the pitch extractor, treatment of the unvoiced and silence frames, etc., it can be concluded that they employ one of these four basic mechanisms in voiced frames: exploitation of the harmonic structure, comb estimation of the clean signal, tunnelling noise estimation (or anti-comb-filtering) which can be employed for SS (HT) or for mask estimation (as in our proposal) and harmonicity mask estimation.

- The maximum number of noise spectral samples which can be recovered in a noisy voiced frame by means of the pitch are (in ideal conditions) the $N(N_p - 1)/N_p$ tunnelling samples, where N is the frame length and N_p the number of periods of the voiced signal. From this it can be deduced that, in order to estimate noise, it is necessary to add more information about the noise and it is just what tunnelling estimation (HT, FPM-NE or our proposal) does when the noise is interpolated by using theses tunnel samples. It can be concluded that (ideally) this kind of techniques achieve optimum noise estimation based on pitch and employing very little information about the noise (the interpolation model).

- It can be shown that mask estimation by means of both tunnelling noise and harmonicity mechanisms yields similar masks. Taking into account that tunnelling noise is optimum (at least, under certain conditions) and the advantages of MD (as compared to SS), we can conclude that the mask estimation mechanisms based on tunnelling or harmonicity, along with MD recognition, provide a very solid framework for pitch-based recognition of voiced frames and that, in ideal conditions, these can be considered as an optimum mechanisms. The experimental results, employing oracle masks, support this assertion.

- Taking into account the optimum voiced mechanisms and the experimental results with oracle masks (in unvoiced and silence frames), we can conclude that the proposed pitch-based noise estimation technique performs reasonably well (with ideal pitch) because its results are close to the limits of the pitch-based ASR techniques (using the minimal noise information). Besides, these results are not very far from the oracle mask results. In order to reach these oracle results it would be necessary to add more information (about noise or speech) in the mask estimation.
- Some ideas presented in this work, such as employing MD or the main source model to obtain a VAD, can be exploited to recognize whispered speech (without pitch).

2.2. Contributions

The main contributions of this Ph.D. dissertation can be summarized as follows:

- We propose a set of asymmetric windows which are applied to the OSA in order to carry out robust feature extraction with low computational cost [34].
- We propose a clean autocorrelation estimator which employs the pitch and can deal with harmonic (not related with pitch) and white-like noises. This estimator is the sifting estimator [33].
- We propose a VAD and a pitch-based noise estimator from a simplify speech model (main source model) which solves many of the problems of similar techniques [32].
- We study different pitch-based techniques, classify them, show their equivalences and point out the limits of the pitch-based recognition, showing that the proposed pitch-based noise estimation technique is close to these limits.

2.3. Future Work

Many of the experiments developed in the Thesis (such as those with ideal pitch) point out possible future work. They can be summarized as follows:

- Regarding **asymmetric windows**, robust feature extraction employing windows centered on the mean pitch speaker could be carried out in order to improve performance as experimental results of Sec. 1.4.1 show.

- Regarding **sifting autocorrelation** a dynamic δ could be applied in order to improve the results (experiments with oracle δ show this, Sec. 1.4.2). The idea of sifting could even be extended, in the sense of not deleting only the products around the main diagonal but also those around other diagonals or other table positions more affected by noise.

- Regarding **pitch-based noise estimation** we can say that the main point is to improve the pitch extraction as shown by the ideal pitch results. If this was done, the technique would almost reach the limits of pitch-based techniques as Tab. 1.7 points out (without the necessity of improving the VAD). One solution could be to consider several pitch candidates at each frame, and each candidate could result in a different noise estimation hypothesis. These parallel hypotheses could be evaluated separately by using missing data marginalization and employing the mask derived from a hypothesized noise estimate. The pitch which gave the highest likelihood would be chosen. This is similar to the SFD (speech fragment decoding) idea which uses top-down speech models to resolve bottom-up signal ambiguity.

- Another interesting work which is pointed out by table 1.7 is trying to reach the oracle mask limits mainly at low SNRs. As we have seen, we can not reach these limits only by means of the pitch. The way to do that would be adding more information about the noise (or speech) to the mask estimator. This information could be dynamically updated in time from silence regions.

- Finally, recognition of speech without or even with multiples pitch values (whispered or vocal harmony speech) is a very interesting line as it is discussed in Sec. 1.5.3.

Bibliografía

[1] Aurora-3-Spanish. Aurora-3, aurora project database: Subset of speechdat-car, spanish database. Technical report, ELRA (European Language Resources Association), 2001. 1.4.1

[2] J. Barker, M. Cooke, and D. Ellis. Decoding speech in the presence of other sources. *Speech Communication*, 45:5–25, 2005. 1.3.1, 1.3.2, 1.5.3

[3] J. Barker, M. Cooke, and P. Green. Robust asr based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise. In *Eurospeech*, pages 213–216, 2001. (document), 1.2, 1.3.2, 1.4.3, 1.4.3

[4] J. Barker, L. Josifovski, M. P. Cooke, and P. D. Green. Soft decisions in missing data techniques for robust automatic speech recognition. In *ICSLP*, 2000. 1.4.3

[5] J. Barker, P.Green, and M.P. Cooke. Linking auditory scene analysis and robust asr by missing data techniques. In *WISP Stratford-upon-Avon*, 2001. 1.3.2

[6] J. Beh and H. Ko. A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech. In *Proc. IEEE ICASSP*, volume 1, pages 648–651, 2003. 1.3.1

[7] A. Bernard and A. Alwan. Low-bitrate distributed speech recognition for packet-based and wireless communication. *IEEE Trans. on Speech and Audio Processing*, 10(8):570–579, 2002. 1.3.1

[8] Herve Bourlard and Stephane Dupont. A new asr approach based on independent processing and recombination of partial frequency bands. In *ICSLP*, 1996. 1.3.1

[9] Guy Brown and Martin Cooke. Computational auditory scene analysis. *Comput. Speech. Lang.*, 8 (4):297–336, 1994. 1.3.2

[10] L. Buera, J. Droppo, and A. Acero. Speech enhancement using a pitch predictive mode. In *ICASSP*, 2008. 1.3.2, 1.3.2, 1.5.1

[11] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34:267–285, 2001. 1.3.1, 1.4.3

[12] C. J. Darwin. Perceptual grouping of speech components differing in fundamental frequency and onset-time. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 33 (2):185–207, 1981. 1

[13] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio. Histogram equalization of speech representation for robust speech recognition. *IEEE Trans. Speech Audio Process*, 13:355–366, 2005. 1.3.1

[14] J. Droppo, L. Deng, and A. Acero. Evaluation of the splice algorithm on the aurora2 database. In *EUROSPEECH*, 2001. 1.3.1

[15] D. Ealey, H. Kelleher, and D. Pearce. Harmonic tunnelling: tracking non-stationary noises during speech. In *EUROSPEECH*, pages 437–440, 2001. 1.3.2, 1.4.3, 1.5.1, 2.1

[16] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32 (6):1109–1121, 1984. 1.3.1

[17] Ronald H. Frazier, Siamak Samsamt, Louis D. Braida, and Alan V. Oppenheim. Enhancement of speech by adaptive filtering. In *ICASSP*, 1976. 1.3.2, 2.1

[18] M. Gales and S. Young. Robust continuous speech recognition using parallel model combination. *IEEE T. Speech. Audi. P.*, 4 (5):352–359, 1996. 1.3.1

[19] J. A. González, A. M. Peinado, A. M. Gomez, J. L. Carmona, and J. A. Morales-Cordovilla. Efficient vq-based mmse estimation for robust speech recognition. In *ICASSP*, 2010. 1.3.1

[20] L. Gu and K. Rose. Perceptual harmonic cepstral coefficients for speech recognition in noisy environment. In *ICASSP*, 2001. 1.5.1

[21] J. Hernando and C.Ñadeu. Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition. *IEEE Transactions on Speech and Audio Processing*, 5 (1):80–84, 1997. 1.4.1

[22] X. Huang, A. Acero, and H. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. 2001. 1.1.1

[23] T. Itoh, K. Takeda, and F. Itakura. Analysis and recognition of whispered speech. *Speech Communication*, 45:139–152, 2005. 1.5.3

[24] Hyoung Gook Kim, Markus Schwab, Nicolas Moreau, and Thomas Sikora. Speech enhancement of noisy speech using log-spectral amplitude estimator and harmonic tunneling. In *Structure*, 2003. 1.3.1

[25] Y. Kuroiwa and T. Shimamura. An improvement of lpc based on noise reduction using pitch synchronous addition. In *IEEE Int. Symp. Circuits and Systems*, volume 3, pages 122–125, 1999. 1.4.2

[26] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Comput. Speech. Lang.*, 9:171–185, 1995. 1.3.1

[27] N. Ma, J. Barker, H. Christensen, and P. Green. Distant microphone speech recognition in a noisy indoor environment: combining soft missing data and speech fragment decoding. In *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, 2010. 1.4.3

[28] N. Ma, P. Green, J. Barker, and A. Coy. Exploiting correlogram structure for robust speech recognition with multiple speech sources. *Speech Communication*, 49:874–891, 2007. 1.3.2, 1.5.1

[29] D. Macho and Yan Ming Cheng. Snr-dependent waveform processing for improving the robustness of asr front-end. In *ICASSP*, 2001. 1.3.1, 1.5.1

[30] D. Mansour and B.H. Juang. The short-time modified coherence representation and noisy speech recognition,. *IEEE Trans. Audio Speech and Signal Processing*, 37:795–804, 1989. 1.4.1

[31] S. L. Marple. *Digital Spectral Analysis with Applications*. Prentice Hall. New Jersey, 1987. 1.4.1

[32] Juan A. Morales-Cordovilla, Ning Ma, Victoria Sánchez, Jose L. Carmona, Antonio M. Peinado, and Jon Barker. A pitch based noise estimation technique for robust speech recognition with missing data. In IEEE, editor, *ICASSP (International Conference on Acoustic, Speech and Signal Processing)*, pages 4808–4811, Mayo, 22-27 2011. 1.4.3, 2.2

[33] Juan A. Morales-Cordovilla, Antonio M. Peinado, Victoria Sánchez, and José A. Gonzalez. Feature extraction based on pitch-synchronous averaging for robust speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 19(3):640–651, Marzo 2011. 1.4.2, 1.4.2, 1.4.2, 1.4.2, 1.4.2, 2.2

[34] Juan A. Morales-Cordovilla, Victoria Sánchez, Antonio M. Peinado, and Ángel. Gómez. On the use of asymmetric windows for robust speech recognition. *Circuits, Systems and Signal Processing (Springer)*, 2011, Abril (aceptado con cambios). 1.4.1, 2.2

[35] P. Moreno. *Speech Recognition in Noisy Environments*. PhD thesis, Carnegie Mellon University, 1996. 1.3.1

[36] Kuldip K. Paliwal and Yoshinori Sagisaka. Cyclic autocorrelation-based linear prediction analysis of speech. In *EUROSPEECH*, 1997. 1.4.1

[37] D. Pearce and H. G. Hirsch. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ICSLP*, volume 4, pages 29–32, 2000. 1.4.1

[38] Antonio M. Peinado and Jose C. Segura. *Speech Recognition over Digital Channels*. Wiley, 2006. (document), 1.1.1, 1.1, 1.3.1, 1.4.3

[39] C. Ris and S. Dupont. Assessing local noise level estimation methods: application to noise robust asr. *Speech Communication*, 34 (2):141–158, 2001. 1.3.2, 1.5.1

[40] Robert Rozman and Dusan M. Kodek. Using asymmetric windows in automatic speech recognition. *Speech Communication*, 2007. 1.4.1

[41] J. Ryalls. *A basic introduction to speech perception*. Speech Science Series, 1997. 1.4.3

[42] M. Seltzer, J. Droppo, and A. Acero. A harmonic-model based front end for robust speech recognition. In *EUROSPEECH*, 2003. 1.3.2, 1.4.2, 1.5.1

[43] B. Shannon and K. K. Paliwal. Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition. *Speech Communication*, 48, no. 1:1458–1485, 2006. 1.3.2, 1.4.1, 1.4.1, 1.4.2, 1.5.1, 2.1

[44] Y. H. Suk, S. H. Choi, and H. S. Lee. Cepstrum third-order normalisation method for noisy speech recognition. *IEE Electronic Letters*, 35(7):527–528, 1999. 1.3.1

[45] v1.1.1 ES 202 050. *Advanced front-end feature extraction algorithm.* ETSI, 2002. 1.3.1, 1.4.2, 1.4.3

[46] v1.1.1. ES 202 211. *Distributed speech recognition; Extended front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm.* ETSI, July 2001. 1.4.2

[47] v1.1.3 ES 201 108. *Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithms.* ETSI, April 2003. 1.3.1, 1.4.1, 1.4.1, 1.4.3

[48] DeLiang Wang and Guy. J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.* IEEE Press/Wiley-Interscience, 2006. 1.1.1, 1, 3, 1.4.3

[49] S. J. Wenndt, E. J. Cupples, and R. M. Floyd. A study on the classification of whispered and normally phonated speech. In *ICSLP, Denver*, 2002. 1.5.3