Circuits, Systems and Signal Processing manuscript No.
(will be inserted by the editor)

On the use of asymmetric windows for robust speech recognition

Juan A. Morales-Cordovilla · Victoria Sánchez · Antonio M. Peinado · Angel M. Gómez

Received: date / Accepted: date

Abstract This paper deals with the problem of searching for a suitable window for robust speech recognition in noisy conditions. A set of asymmetric windows, so-called $DDR_{c,w}$, are proposed which are controlled by two parameters, center c and width w. These windows act over the OSA (One-Sided Autocorrelation) in order to perform spectral estimation. The two parameters, c and w, allow us to control the level of weight given to the first noisy autocorrelation coefficients and to emphasize the important ones. Finally, it is shown that the best window of the proposed set is the $DDR_{62,200}$. This window is centered around the average pitch of human speech and it provides a higher speech recognition performance over the Aurora-2 and Aurora-3 databases than those obtained by previously proposed windows.

Keywords Robust speech recognition · feature extraction · autocorrelation · OSA · HASE · DDR · asymmetric window · pitch · gender recognition

1 Introduction

Acoustic noise represents one of the major challenges for automatic speech recognition systems. Many different approaches have been proposed to deal with this problem [1]. In this paper we are interested in improving feature extraction without any prior knowledge about the noise. When performing feature extraction, one of the first steps is usually to determine the PSD (Power Spectral Density) of the speech signal.

There are two basic ways to obtain the PSD of a speech signal [5]. One way is by means of parametric methods, such as the classical LPC spectrum or those based on all-pole modelling of the causal part of the autocorrelation sequence [2], [3]. The other way is by means of non parametric methods. In

CSSP 2011 v.. The authors are with the department of Teoría de la Señal, Telemática y Comunicaciones and CITIC-UGR, Universidad de Granada, 18071-Granada, Spain
E-mail: {jamc,amp,victoria,amgg}@ugr.es

this case the PSD can be directly obtained either from the signal or from the signal autocorrelation. The cepstral coefficients obtained from the PSD of the signal after filtering by a Mel filter bank are called MFCC (Mel Frequency Cepstral Coefficients) meanwhile those obtained from the autocorrelation are called AMFCC (Autocorrelation MFCC) [4].

In both cases, before applying the Fourier Transform, it is necessary to apply a suitable analysis window. Many different windows have been proposed in order to perform a good estimation of the signal spectrum. Each of them has its advantages and disadvantages. Although symmetric windows have been traditionally employed, asymmetric windows have also been proposed [6]. A recent work [4] proposes to estimate the spectrum by means of the HASE (Higher-lag Autocorrelation Spectrum Estimation) technique. It involves estimating the signal spectrum by means of the One-Sided (causal part) Autocorrelation (OSA). In this method lower-lag autocorrelation coefficients are discarded (considering that broadband noise distortion affects mainly those ones) and only higher-lag ($>$ 2ms) autocorrelation coefficients are used. This is done by applying an specially designed window function, the Double Dynamic Range (DDR) Hamming window, to the one sided autocorrelation sequence. The HASE method outperforms the aforementioned methods in noisy conditions.

The windowing process of the HASE method can be also seen as an asymmetric weighting of the different autocorrelation coefficients of the OSA. Under this point of view, the HASE technique uses a window which gives a null weight to the first autocorrelation coefficients and the rest of the coefficients are weighted according to a DDR window. Following this idea of weighting factors which act on the OSA, in this paper we will search for a window (or a set of weights) that increases AMFCC robustness against noise.

The structure of the paper is as follows. In section 2 a set of windows, inspired by the HASE method, is proposed. Next, in section 3, it is shown that, for voiced frames, the most robust windows are those centered around the pitch. In section 4 the best window of this set is found, comparing its effectiveness with similar windows. Finally in section 5, the most important conclusions will be summarized.

2 Window parametrization

Within the framework of AMFCC feature extraction, we want to find a suitable window which, when applied to the OSA, improves robustness against noise. We will assume here a classical speech recognition system, in which the models are trained with clean speech feature vectors and, to avoid mismatch, the window used for training is the same one as that used for test. A good window could be one which minimizes the distance between noisy and clean vectors. In order to carry out the search for this window, we will characterize the set of possible windows by two parameters.

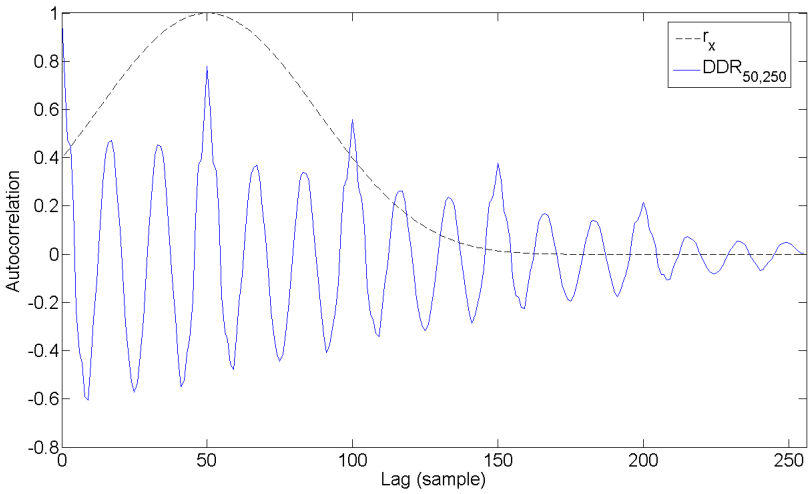

Fig. 1 Example of a $DDR_{50,250}$ window applied to the OSA of a voiced frame with a pitch value of 50 samples.

To build our parametric model, we will start with a DDR Hamming window but controlled by two parameters c and w, resulting in what will be called $DDR_{c,w}$. Parameter w indicates that the proposed window is based on a DDR of width w and parameter c that the center or maximum of the window is located on the c-lag autocorrelation coefficient. Specifically the expression for the proposed $DDR_{c,w}$ window of length L would be given as:

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

where DDR_w is a DDR window of width w that is obtained autocorrelating a Hamming window of width $w/2$ [4]. According to this construction, the HASE window proposed by Shannon in [4] is equivalent to a $DDR_{135,240}$ for a speech signal sampled at 8 kHz. Fig. 1 shows a $DDR_{50,250}$ of length $L = 256$ together with the first 256 coefficients of the OSA of a voiced frame.

This sort of parametrization lets us vary (through parameters c and w) the weights given to the first autocorrelation coefficients (usually more contaminated by noise) without the need of discarding them completely (as it is done in HASE) since these first coefficients could carry useful information. It also allows the possibility of deciding which are the most important coefficients and, consequently, placing the center of the window on them. In section 3, it will be shown that the best place to locate the center is just over the pitch or its corresponding multiples.

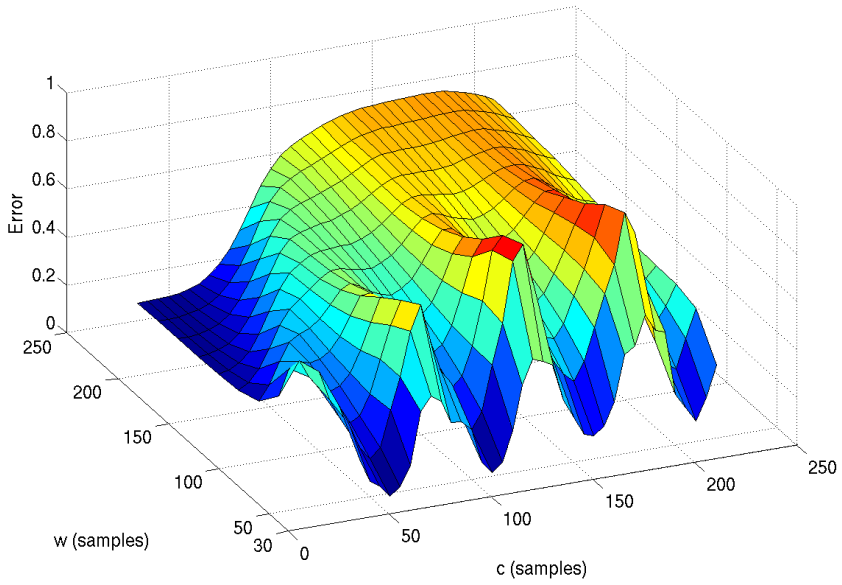

Fig. 2 Cepstral error, *Err(c,w)*, surface for a voiced sound (pitch=50 samples) contaminated with white noise as a function of the center (c) and width (w) of the analysis window $DDR_{c,w}$.

3 Window for voiced signal

In order to study which $DDR_{c,w}$ window is the best one for voiced signals, a frame of clean voiced speech signal has been contaminated with 100 instances of AWGN noise to achieve an SNR of 0dB. In particular, this frame has been extracted from a recording of the vocal 'e' with a 50-sample pitch. Its OSA is shown in Fig. 1. The error surface obtained when a $DDR_{c,w}$ window is applied is plotted in Fig. 2. This error has been computed as the averaged distance between the clean and noisy AMFCC cepstrum as,

$$\mathrm{Err(c,w)} = \frac{1}{N}\sum_{n=1}^{N} ||C_x^{c,w} - C_{y_n}^{c,w}|| \tag{2}$$

where $C_x^{c,w}$ is the clean AMFCC, $C_{y_n}^{c,w}$ the nth noisy instance AMFCC and N is the number of noise instances (100 in our experiment).

It can be observed that several deep valleys appear, located at $c = 50, 100, 150, \ldots$ etc. From this fact we can draw the following conclusion: the minimum cepstral error is reached when the $DDR_{c,w}$ window has its center around the pitch of the clean signal or its multiples. This conclusion has been drawn for white noise but the results obtained in the next section validate it for other types of noise.

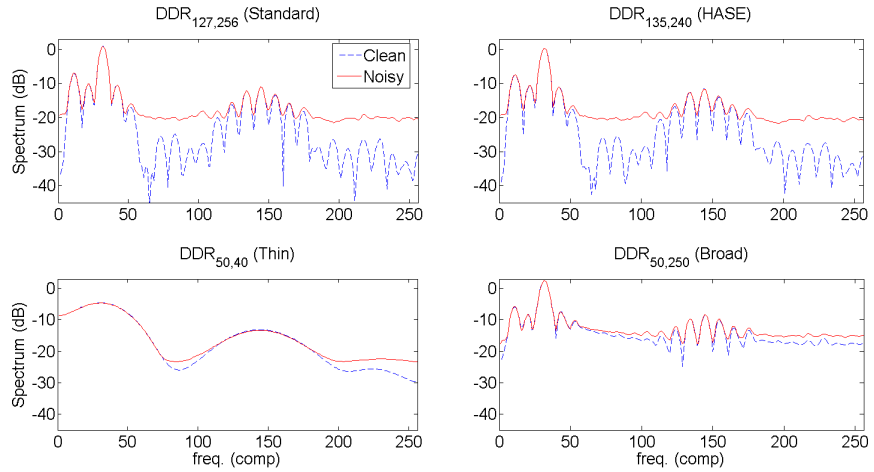

Fig. 3 Averaged spectra of four different windows applied to a vocal with pitch=50 samples contaminated with white noise.

Lets analyze now what happens in the spectral domain. Fig. 3 shows the clean spectrum and the averaged noisy spectra for four different windows: $DDR_{127,256}$(Standard), $DDR_{135,240}$(Shannon), $DDR_{50,40}$(Thin) and $DDR_{50,250}$(Broad).

Although windows centered on the pitch with high w values, such as $DDR_{50,250}$, are short of dynamic range (i.e. the window has not enough spectral range to cover the 80 dB necessary for the speech autocorrelation), this does not constitute a serious problem since it only produces a poor characterization of the spectral valleys. The reason for this short dynamic range is due to the truncation at lag 0 (in contrast to the Shannon and Standard windows) so we can consider those windows as DDR windows with a superposed rectangular window. As the rectangular window has a short dynamic range, this effect translates to the resulting windows (see Fig. 3 for the case of $DDR_{50,250}$). However, spectral valleys, which are between the formants, are not as important for speech recognition as spectral peaks [7]. The most important thing is to have a good characterization of the formants and, in fact, this is what windows centered on the pitch do. This is so because the information regarding the spectral envelope (and the formants) is located on the zero lag area as well as on the pitch and on integer lags multiple of the pitch. In fact, as it will be shown in section 4.2, the smaller dynamic range of the proposed windows barely affects recognition performance in clean conditions. An additional advantage of the proposed windows is that they reduce the mismatch between training and testing conditions (models are trained in clean condition with the same window) as can be observed in figure 3, a fact that contributes to the improvement of the speech recognition system performance.

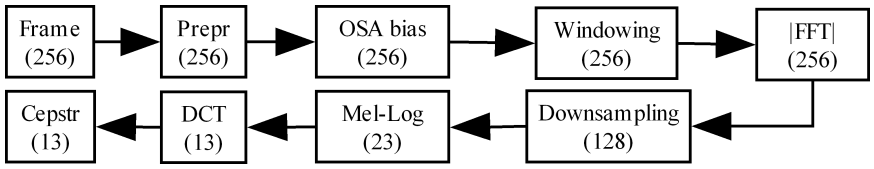

Fig. 4 Processing path of our front end, from a speech frame (with 256 samples) till the final cepstral vector (13 coefficients).

4 Experimental window

4.1 Experimental framework

Experiments are carried out with the connected digit databases Aurora-2 and Aurora-3 both sampled at 8 kHz. Aurora-2 contains utterances contaminated at 20, 15, 10, 5, 0 and -5 dB with ten different additive noises: subway, babble, car and exhibition for Set-A, restaurant, street, airport and train for Set-B, and convolutional subway and street noises for Set-C. Aurora-3 is a database with real noise contaminated speech. This database contains in-car speech recorded by several microphones placed at different places, what provides 3 different conditions: well-matched (WM), medium mismatch (MM) and high mismatch (HM). Both databases will be used in their classical configuration: each digit is modeled with an HMM of 16 states and 3 Gaussian per state except silence that has 3 states and 6 Gaussians per state. More details about these databases and the back-end configurations employed can be found in [9] and [10].

The feature extractor used is very similar to that from ETSI [8]. The main difference is that we extract AMFCCs instead of MFCCs. The feature extraction process is depicted in Fig. 4. The number that appears at each stage in Fig. 4 indicates the number of coefficients resulting from every stage. We comment next the most relevant details. Just as done in [4], the windows are applied on the biased OSA (256 coefficients). After applying the FFT and calculating its magnitude, the resulting 256 coefficients are downsampled to obtain 128 coefficients that are passed to the Mel filter bank. Here on, the rest of the feature extraction procedure coincides with that from ETSI, finally obtaining a cepstral vector with 13 MFCC coefficients (C0, C1,...,C12). The final feature vector is formed by the 13 static MFCCs plus their corresponding delta and delta-delta coefficients (a vector of 39 coefficients in total) . Every vector is also compensated by Cepstral Mean Normalization (CMN).

Finally, it is worth mentioning that we always use the same window for both, training and testing stages and, in the case of Aurora-2, training is carried out only with clean speech. In addition, although the proposed windows have been developed considering their application to voiced frames, the same window will also be applied to unvoiced and silence frames. For unvoiced frames this could involve a loss of information (their spectral information is mainly contained at the first lags) but, if training and testing are done with

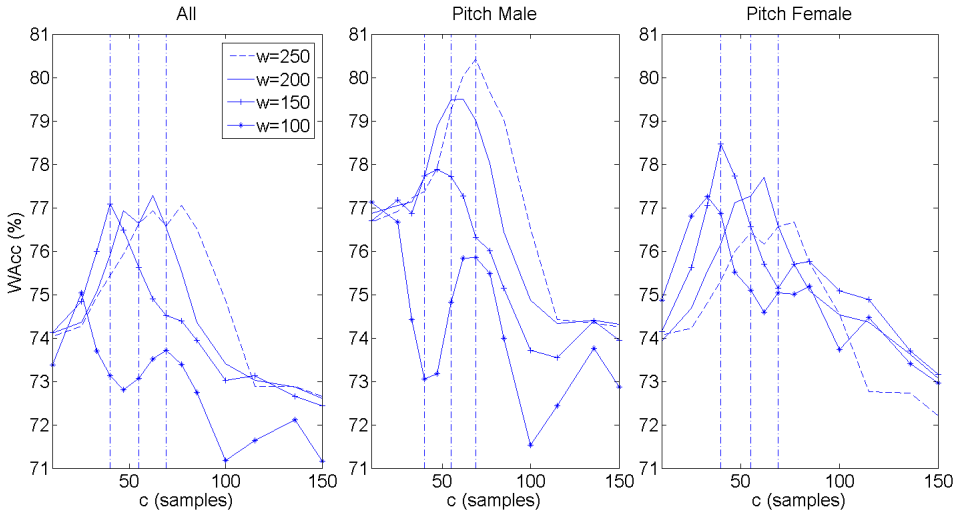

Fig. 5 Averaged recognition accuracy WAcc(%) when considering different c and w values, for the complete Aurora-2 databases (*All*), only with male speech (*Pitch Male*) and only with female speech (*Pitch Female*). Mean pitch of each group is indicated with dashed vertical lines at 55, 62 and 40 samples respectively

.

the same window, it does not affect recognition performance much, as we will see in the clean condition experiments in the following section. In the case of silence frames, there is no information to loose so the use of our asymmetric windows will be always beneficial.

4.2 Experimental results

Fig. 5 shows the averaged word accuracy over Aurora-2 (WAcc (%) along the A, B and C sets, from 0 to 20 dB). In the left figure all the utterances have been considered, in the center figure only the male utterances are included (those with pitch higher than 55 samples), and in the rigth figure only the female ones (pitch lower than 55 samples). Each result has been obtained using different center (c) and width (w) values for the $DDR_{c,w}$ analysis window.

If we consider all the sentences, the best recognition results are obtained by windows with center around 60 samples and width of around 200 samples. It is worth mentioning that 55 samples is the mean pitch for Aurora-2 (40 samples or 200 Hz for female speech and 69 samples or 116 Hz for male speech), while the maximum in that figure corresponds to the window $DDR_{62,200}$ (77.28%). Comparing this result with that achived by means of a HASE window ($DDR_{135,240}$ - 72.88%), we can observe a significant improvement. In case of considering only the male sentences, the best recognition results are obtained for a window centered on the mean pitch (69 samples)

Window	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Average (20-0 dB)
Hamming (FE)	99.07	97.21	92.90	78.37	47.04	24.05	13.57	67.91
$DDR_{135,240}$ (HASE)	99.16	97.37	94.32	84.55	59.42	28.72	15.08	72.88
$DDR_{55,200}$ (Mean pitch)	98.91	96.32	93.43	85.85	68.70	38.93	17.43	76.65
$DDR_{62,200}$ (Opt. Aurora-2)	99.05	96.87	94.31	87.08	69.93	38.21	17.09	77.28

Table 1 Word accuracies obtained by different windows tested with Aurora-2 (Set A, B and C) for different SNR values.

and width 250, while when we have only the female sentences the best results are obtained for a window centered on the mean pitch (40 samples) and width 150. All of this reinforces our hypothesis that the highest robustness against noise is reached when the $DDR_{c,w}$ window is centered around the pitch value as here the local SNR is higher. Another important issue has to do with the width of the DDR window. It should be wide enough to cover the different pitch values, but not too much, because it could overweight the first autocorrelation coefficients and then reduce recognition accuracy. This is shown by the fact that the width of the best window (male $DDR_{69,250}$, all $DDR_{62,200}$ and female $DDR_{40,150}$) decreases as the center of the window moves to the left.

Table 1 shows word accuracies obtained by four selected analysis windows tested over Aurora-2 (Set A, B and C) for different SNR values, namely Hamming, $DDR_{135,240}$, $DDR_{55,200}$ and $DDR_{62,200}$. The Hamming window is directly applied on the signal (not on the OSA) obtaining MFCCs as in the ETSI feature extractor. $DDR_{135,240}$ is applied on the OSA, it obtains the same AMFCCs as in HASE method. $DDR_{55,200}$ is a $DDR_{c,w}$ window centered on the mean pitch (55 samples) and $DDR_{62,200}$ is the best window for the whole Aurora-2 database. The last two are the windows proposed by us and, as can be seen, they are centered on or close to the mean pitch.

Results show that our two windows represent an improvement over the Hamming and HASE windows. Our best window for Aurora-2 ($DDR_{62,200}$) improves in almost a 4.5 % in comparision with the HASE result. The results for clean condition are also good in spite of the fact that our windows have a small dynamic range and there is certain information loss in unvoiced frames. This confirms our hypothesis mentioned at the end of Sec. 4.1.

Results with Aurora-3 and the same windows are depicted in table 2. It can be observed that the two proposed windows again improve the results of HASE, mainly for the worst condition (high mismatch). For this case, $DDR_{55,200}$ and $DDR_{62,200}$ improves a 3.76 % and 2.5 % the HASE result, respectively.

Taking all of this into account, we could finally consider the proposed $DDR_{62,200}$ window as the optimum window since it gives good results for both, Aurora-2 and Aurora-3.

Window	WM	MM	HM	Average
Hamming (FE)	89.08	82.15	64.51	78.58
$DDR_{135,240}$ (HASE)	89.76	83.16	76.39	83.10
$DDR_{55,200}$ (Mean pitch)	89.85	82.87	80.15	84.29
$DDR_{62,200}$ (Opt. Aurora-2)	90.89	84.22	78.89	84.67

Table 2 Word accuracies obtained by the different windows applied to Aurora-3 Spanish (real noise) under well-matched (WM), medium (MM) and high mismatch (HM) conditions.

5 Conclusions

In this paper the problem of searching for a suitable window for robust speech recognition has been addressed. Inspired by the HASE technique we have proposed a set of windows called $DDR_{c,w}$ controlled by two parameters (center c and width w) that perform an asymmetrical windowing or weighting of the autocorrelation coefficients of the OSA. These two parameters allow us to control the level of weight given to the first noisy autocorrelation coefficients.

It has been shown that the highest robustness, for a voiced signal, is obtained by windows centered around the pitch values as, in this case, formant information is better preserved. This has been confirmed by recognition experiments based on the speaker gender. For both groups (male/female), the best results were obtained when the analysis windows were centered on gender average pitch.

Finally, two windows have been proposed ($DDR_{55,200}$ and $DDR_{62,200}$) and evaluated over Aurora-2 and Aurora-3 databases. Both windows have obtained better overall results than those obtained by the Hamming window (as used in the ETSI Front End [8]) and HASE over Aurora-2 and Aurora-3 test sets.

Acknowledgment

This work has been supported by the Spanish MEC/FEDER project TEC2010-18009 and project CEI BioTIC GENIL (CEB09-0010).

References

1. A. M. Peinado and J. C. Segura, *Speech Recognition over Digital Channels*, Wiley, July 2006.
2. D. Mansour and B.H. Juang, "The Short-Time Modified Coherence Representation and Noisy Speech Recognition," *IEEE Trans. Audio Speech and Signal Processing*, vol. 37, pp. 795-804, 1989.
3. J. Hernando and C. Nadeu, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp.80-84, 1997.

4. B. Shannon and K.K. Paliwal, "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition," *Speech Communication*, vol. 48, no. 1, pp. 1458-1485, Jan. 2006.
5. John G. Proakis y Dimitris G. Manolakis. "Tratamiento digital de señales," *Ed. 3 of Prentice Hall*, 2000.
6. Robert Rozman, Dusan M. Kodek. "Using asymmetric windows in automatic speech recognition," *Speech Communication*, Jan. 2007.
7. Douglas O'Shaughnessy "Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, pp. 2965-2979, May 2006.
8. ETSI ES 201 108 v1.1.3. Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithms. April 2003.
9. D. Pearce and H. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. of ICSLP'2000*, vol. 4, pp. 29-32, October 2000.
10. Aurora Project Database: Subset of SpeechDat-Car - Spanish Database. *European Language Resources Association (ELRA)*, 2001.