# WIGNER DISTRIBUTION - A NEW METHOD FOR HIGH-RESOLUTION TIME-FREQUENCY ANALYSIS OF SPEECH SIGNALS

W. Wokurek, G. Kubin, F. Hlawatsch

Institut für Nachrichten- und Hochfrequenztechnik, TU Wien
Gusshausstr. 25/389, A-1040 Vienna, Austria

## ABSTRACT

Two methods for the time-frequency analysis of speech signals are compared: the tradionally used Spectrogram and the Smoothed Pseudo Wigner Distribution (SPWD). It is shown that the time and frequency resolutions of the Spectrogram are restricted by the uncertainty relation while SPWD allows arbitrarily high resolutions. If the analysis parameters are chosen carefully SPWD yields more accurate signal representations than the Spectrogram. This is exemplified by a "microscopic" analysis of vowels and unvoiced stop consonants.

## 1. INTRODUCTION

The Wigner Distribution (WD) is a method for the time-frequency analysis of signals. Along with the Spectrogram the WD is a member of a special class of bilinear, shift- invariant signal representations (Cohen class [1] p.376). Within this paper we compare a somewhat modified WD, i.e. the Smoothed Pseudo Wigner Distribution (SPWD) to the Spectrogram, first with respect to the basic features of time- and frequency resolutions (Sections 2-4). In sections 5 and 6 we compare the results of representing speech signals through SPWD and Spectrogram.

Observing the fact that SPWD enables high-resolution signal representation, we analyze short speech segments of a few pitch periods of length. Therefore it is pointless to compare the SPWD to Spectrograms of high frequency resolution (45 Hz) because they do not display the fine-structure in time that we will see in the SPWD. As a compromise between Spectrograms of high frequency resolution (which do not show the time- structure) and those of high time resolution (which smear out the formant structure etc.) we find the spectrogram of 300 Hz frequency resolution as a suitable partner for the comparison with the SPWD of vowels (see section 5). In the case of unvoiced stop consonants, equal frequency resolution of the Spectrogram and SPWD is chosen for the analysis of the whole explosion interval of several centiseconds duration because there is no significant time structure of the noise-like excitation observed (section 6).

## 2. DISTORTION OF TIME-FREQUENCY ANALYSIS DUE TO LIMITED RESOLUTION

The aim of a time-frequency representation of any signal is to show the stucture of the signal and not that of the analysis method. One of the basic distortions of any analysis method is its limited resolution. To study the nature of time resolution consider an impulse in the time domain as shown in Fig.1.

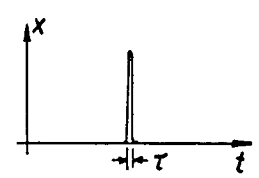


Fig.1: Impulse of length τ in the time domain.

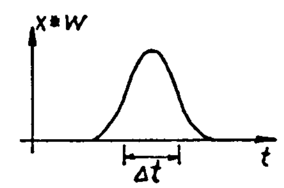


Fig.2: Representation of an impulse with a time resolution of Δt.

If we represent this impulse with a time resolution $\Delta t \gg \tau$ the impulse will be widened to the duration $\Delta t$ (see Fig.2).

The mathematical model of this effect is the convolution of the signal $x$ with a window function $w$ of time width $\Delta t$:

$$[x \underset{t}{*} w](t) = \int_{\mathbb{R}} x(t-\tau)\, w(\tau)\, d\tau \qquad (1)$$

A second interpretation of (1) is lowpass filtering. If the signal contains oscillations of periods less than the time resolution $\Delta t$, these components of the signal will be supressed in the representation.

A similar effect is caused by the frequency resolution $\Delta f$. All signal components will be widened by $\Delta f$ in the frequency direction. On the other hand all signal changes within a frequency range of $\Delta f$ will be canceled.

## 3. COUPLING OF THE RESOLUTIONS OF THE SPECTROGRAM

The Spectrogram is defined as the square magnitude of the Short-Time Fourier Transform (STFT). The signal is multipied by a window $w(\tau)$ that is shifted to the instant of analysis t. The Fourier Transform of this product is associated with the instant t.

$$S_x(t,f) = \left| \int_{\mathbb{R}} x(\tau)\, w(\tau-t)\, e^{-j2\pi f\tau}\, d\tau \right|^2 \qquad (2)$$



Using elementary signal theory, we can recast eq. (2) in a form containing a convolution with the window $w(t)$

$$S_x(t,f) = \left| [e^{-j2\pi ft}\, x(t)] \underset{t}{*} w(-t) \right|^2 \qquad (3)$$

or with its spectrum $W(f)$,

$$S_x(t,f) = \left| [e^{j2\pi ft}\, X(f)] \underset{f}{*} W(f) \right|^2 \qquad (4)$$

This shows us the simultaneous determination of both the time and the frequency resolution of the Spectrogram by a single window function. Like any other function, the window satisfies the uncertainty relation (5), where c is a constant that depends only on the definitions of $\Delta t$ and $\Delta f$ and is of the order 1.

$$\Delta t . \Delta f \geq c \qquad (5)$$

The uncertainty relation (5) restricts the allowed values of the time and frequency resolutions of the Spectrogram to the region U shown in Fig. 3.

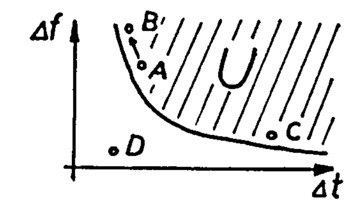


Fig.3: Restriction of the Spectrogram resolutions by the uncertainty relation

Because the product of the Spectrogram resolutions $\Delta t . \Delta f$ cannot be less than the constant c, it is impossible to choose both resolutions arbitrarily high (i.e. $\Delta t$ and $\Delta f$ arbitrarily small) at the same time. This implies the necessity of trading-off between these two resolutions. If the time resolution is increased (smaller $\Delta t$), the Spectrogram must have a poorer frequency resolution (greater $\Delta t$, see Fig.3: movement from point A to B). The dual case is the choice of higher frequency resolution (Fig.3: point C), thus decreasing the time resolution.

## 4. INDEPENDENCE OF THE RESOLUTIONS OF THE SPWD

The Wigner Distribution (WD) of a signal $x(t)$ is defined by (6)

$$WD_x(t,f) = \int_{\mathbb{R}} x(t+\tfrac{\tau}{2})\, x^*(t-\tfrac{\tau}{2})\, e^{-j2\pi f\tau}\, d\tau \qquad (6)$$

and its features are described in [1] extensively. The WD does not show any effect of limited resolution, but in the case of fairly complex signals such as speech the result is quite unreadable owing to the occurence of *interference terms*, described in [2] (see section 5 also). Therefore we consider the SPWD of the signal which is defined as a WD with arbitrary smoothing:

$$SPWD_x = WD_x \underset{t}{*} u(t) \underset{f}{*} v(f) \qquad (7)$$

Smoothing in both the time and frequency direction is performed by two independently chosen arbitrary windows $u(t)$ and $v(f)$, respectively. Because of the independence of the smoothing functions, the resolutions of the SPWD are not restricted by the uncertainty relation (5) (see Fig 3: point D).

Yet, from a practical point of view, the resolutions of the SPWD are restricted by the occurence of *interference terms* and depend on the structure of the signal in that way. The analysis of speech signals shows that with equal frequency resolution (e.g. 100 Hz), the SPWD allows a substantially higher time resolution than the Spectrogram (e.g. 1 ms instead of 10 ms).

An interesting insight into the relation between the Spectrogram and WD is obtained from the following equation:

$$S_x = WD_x \underset{t}{*} \underset{f}{*} WD_w \qquad (8)$$

This equation proves that the Spectrogram is the WD of the signal smoothed in both directions with the WD of the Spectrogram window. In contrast to (7) the time and frequency smoothing is determined by one and the same window $w(t)$ as we have seen already in (3) and (4) and this is why Spectrogram resolutions are bounded by the uncertainty relation (5) ([1] p.382).

## 5. ANALYSIS OF VOWELS BY SPWD AND SPECTROGRAM

Figure 4 shows a *contour plot* of the SPWD of three successive pitch periods extracted from the German vowel [a:] spoken by a male subject. This representation displays the following features:

(1) Quasi-periodic excitation of the vocal tract by wideband narrow-time impulses every 10 msec. The time resolution of approx. 0.5 msec is sufficient to prove that these impulses have a time width of 1 msec or less.
(2) Exponential decay of three formants at the frequencies F1 = 0.7 kHz, F2 = 1.25 kHz, and F3 = 2.6 kHz. The frequency resolution of appprox. 100 Hz is sufficient to separate the individual formants and to measure their bandwidths during the intervals outside the excitation impulses.

(3) Besides these *signal terms* (formants, impulses), the SPWD contains *interference terms*. They are governed by a simple geometrical rule [2], i.e. they always lie half-way between two signal terms and oscillate in the direction perpendicular to the line connecting the two signal trems. These oscillations have a period in the time-frequency plane that is inverse proportional to the distance of the signal terms. The oscillatory nature of interference terms is the key to their suppression in any bilinear time-frequency representation. In SPWD, this is achieved by smoothing with the two independent window functions according to (7). The amount of smoothing must be matched to the signal structure:

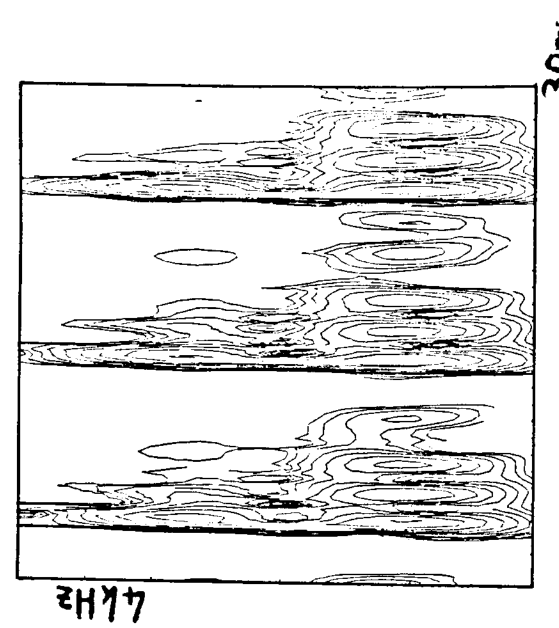


Fig. 6: Spectrogram [a:] from [ta:t] Δf=600Hz, Δt=1ms, ΔtΔf=0.6

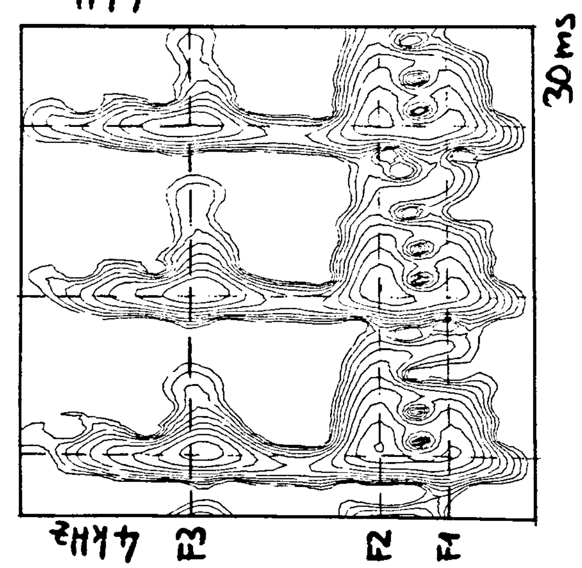


Fig. 5: Spectrogram [a:] from [ta:t] Δf=300Hz, Δt=2ms, ΔtΔf=0.6

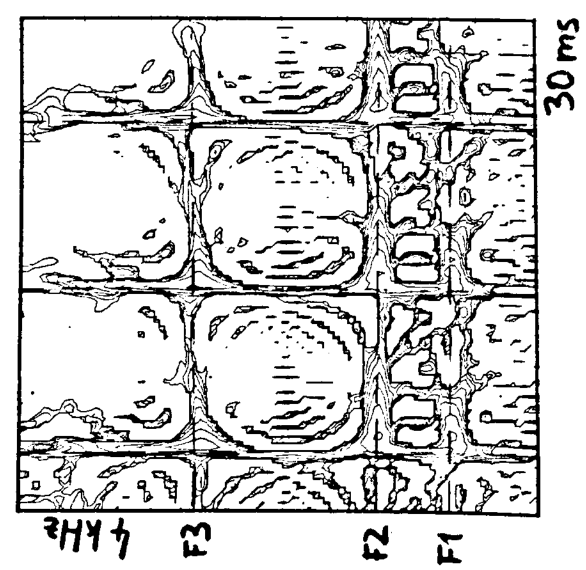


Fig. 4: SPWD [a:] from [ta:t] Δf=100Hz, Δt=0.5ms, ΔtΔf=0.05

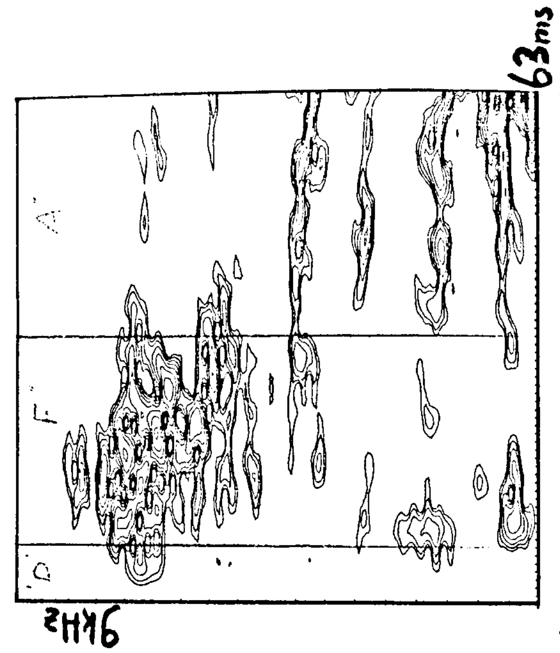


Fig. 9: Spectrogram of explosion interval in first [t] from [ta:t] Δf=100Hz, Δt=6ms, ΔtΔf=0.6

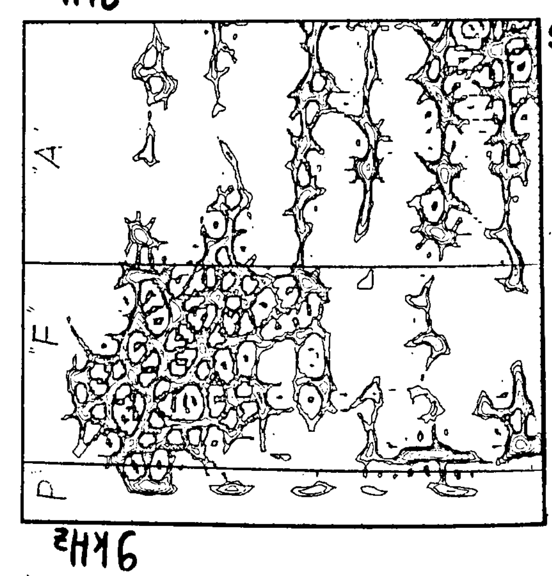


Fig. 8: SPWD of explosion interval in first [t] from [ta:t] Δf=100Hz, Δt=1ms, ΔtΔf=0.1

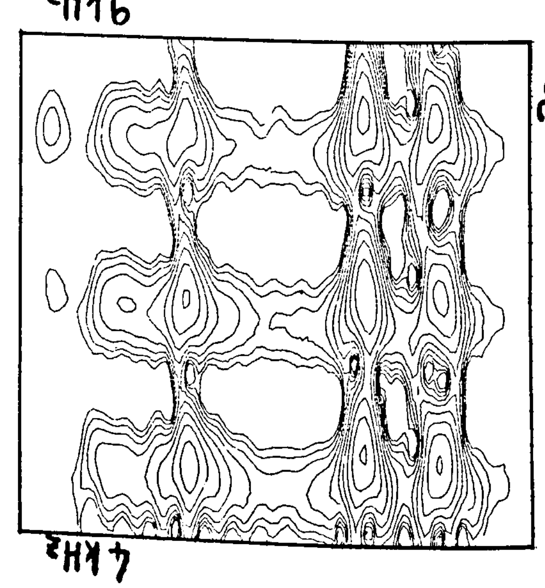


Fig. 7: Spectrogram [a:] from [ta:t] Δf=100Hz, Δt=6ms, ΔtΔf=0.6



The frequency resolution $\Delta f$ = 100 Hz is just great enough to damp interferences between successive pitch periods (remember interference oscillations to occur perpendicular to the line from one excitation impuls to the next, i.e. parallel to the frequency axis!). The time resolution $\Delta t$ = 0.5 msec is great enough to damp most of the interferences between neighbouring formants. Accordingly, oscillations in the time direction can only be observed between F1 and F2, the two formants closest to each other.

Figure 5 shows a Spectrogram of the same signal segment with resolutions $\Delta f$ = 300 Hz and $\Delta t$ = 2 msec. Note that the product of these resolutions equals 300 Hz · 2 msec = 0.6 which is more than ten times the product of resolutions of SPWD in Figure 4 (100 Hz · 0.5 msec = 0.05). The Spectrogram's resolutions are already chosen so as to achieve a signal representation as close as possible to SPWD. A simultaneous improvement of the Spectrogram's resolutions is impossible due to (5). Therefore the Spectrogram evidences much broader excitation impulses in the time dimension as well as much wider formants in the frequency dimension than SPWD. The inherently stronger smoothing of the Spectrogram renders better suppression of interference terms (though they are still perceivable between F1 and F2), yet worse fidelity in signal terms than SPWD. As interference terms are predictable from the above geometrical rule, SPWD is better suited to the analysis of vowels than the Spectrogram.

One may conjecture that a change of the Spectrogram window function $w(\tau)$ in (2) may improve its resolutions. This has to be refuted when studying Figures 6 and 7. In Figure 6, the time resolution of the Spectrogram is improved to $\Delta t$ = 1 msec, thus approaching the value of $\Delta t$ for SPWD in Figure 4. Due to the uncertainty relation (5), the frequency resolution goes up to $\Delta f$ = 600 Hz so that the two lower formants F1 and F2 are merged into a single unstructured lump stretching over several hundred Hz. In Figure 7, the frequency resolution of the Spectrogram is improved to $\Delta f$ = 100 Hz as is the case for SPWD in Figure 4. As time resolution has to go up to 6 msec, excitation impulses are broadened drastically and spilled over the formant structure even into the interval of the pitch period without glottal excitation. Therefore, formant bandwidth measurements are again more difficult than with SPWD, inspite of the high frequency resolution of Figure 7.

Summarizing we observe that the Spectrogram is not suited for *simultaneous* display of both the excitation and the formant structure of vowels whereas SPWD has this property notwithstanding its (easily controlled) interference terms.

## 6. ANALYSIS OF UNVOICED STOP CONSONANTS BY SPWD AND SPECTROGRAM

Figures 8 and 9 show the *explosion* interval (60 msec) of the first [t] in the German word [ta:t] making use of SPWD and Spectrogram, respectively. For the sake of comparison, both displays have a frequency resolution of 100 Hz and associated time resolutions of 1 msec (SPWD) and 6 msec (Spectrogram). The explosion interval consists of three more or less separable phases:

1. An impulse-like transient (about 4 msec) due to the release of the pressure built up behind the vocal-tract closure (*plosion* phase P).
2. A noise phase extending from approx. 4 kHz to 8 kHz (25 msec) due to the turbulent air flow at the opening constriction (*frication* phase F).
3. A noise phase with a formant structure (30 msec) due to the resonances of the open vocal tract excited by turbulent air flow at the glottis (*aspiration* phase A).

The adavantges of SPWD over the Spectrogram for the analysis of this type of sounds can be summarized as follows:

(1) The short impulse of the plosion phase P is readily seen in SPWD whereas the Spectrogram is not able to resolve this temporal fine structure (at the given frequency resolution).
(2) The boundary between frication phase F and aspiration phase A is more pronounced in SPWD than in the Spectrogram.
(3) Noise-like excitation of the vocal tract manifests itself as a very specific meshy *texture* in SPWD which is clearly distinguishable from deterministic excitation as seen in Figure 4. With the Spectrogram, noise-like excitation induces no significant changes in the texture if the contour plots when compared to deterministic excitation as seen in Figures 5, 6, and 7.

## 7. CONCLUSIONS

From the above discussion, it should be clear that SPWD is superior to the Spectrogram for the time-frequency analysis of speech signals as typified by the examples given in sections 5 and 6. It should be kept in mind, however, that the comparison was made on the basis of very short signal segments so as to emphasize SPWD's character as a time-frequency "microscope". If the analysis interval is extended to 1 second or more both the resolutions of video displays and the human eye become insufficient to realize the differences of the two methods. Anyway, these long-time displays are only useful for the compressed visualization of slowly time-varying and global features characterizing whole syllables or words. For the detailed high-resolution study of rapidly time-varying speech phenomena, preference is to be given to the new method.

## 8. REFERENCES

[1] T. Claasen, W. Mecklenbräuker, 1980; The Wigner Distribution - a Tool for the Time-Frequency Signal Analysis; Philips J. Res. Vol. 35 pp. 217-250, 276-300, 372-389

[2] F. Hlawatsch, 1984; Interference Terms in the Wigner Distribution; International Conference on Digital Signal Processing; Florence, Italy