Adaptive Algorithm For Detecting And Reducing Sibilants In Recorded Speech

Adaptive Algorithm For Detecting And Reducing Sibilants In Recorded Speech

Martin Wolters*, Markus Sapp, Jörg Becker-Schweitzer Institute of Communication Engineering (IENT), Aachen University of Technology (RWTH), Germany (*now with Telos Systems, Cleveland, USA)

Abstract

Different concepts for reducing sibilants in recorded speech have been developed, and have been successfully implemented as 'De-Esser' hardware. Most of them are working in the analog domain, only a few implementations using adaptive mechanisms. Therefore a digital algorithm is introduced which basically uses psychoacoustic and physical units to detect sibilancy and to adapt a time-variant bandpass filter to perform the de-esser operation.

Overview

Two examples which demonstrate the need for an adaptive DeEsser algorithm are the broadcast studio and the auditorium. In both situations there may be several different announcers but no possibility to change parameters between each case by hand, either because there is no technician available or the needed time is missing. An adaptive algorithm might be helpful in more sophisticated situations as well, e.g. mastering, where the influence on the signal should be as small as possible and where the algorithm should be adjusted to the current signal very precisely.

To classify the different types of sibilants new investigations on these sounds in German, English, Spanish and French have been made. Connecting with phonetic studies, statistical evaluations of a listening test led to a detailed description of the common and differing properties in the time- and frequency-domain. [Wol97]

To find an adaptive algorithm which considers these properties, a reliable method is needed for detecting the different kinds of sibilants. In [Sap98] it is shown that the psychoacoustic unit 'sharpness' is well correlated with the appearance of these sounds (see figure 1) and that the critical-band intensities - an intermediate step in calculating sharpness - can be utilised to get information about the spectral properties of each sibilant.

This paper focuses on an effective real-time implementation of an adaptive DeEsser algorithm on digital signal processors. Because of the complexity in calculating sharpness a two-step-detector is introduced. The first step uses zero-crossing rate and energy of the signals to roughly detect possible sibilants. But these physical units do not provide information about the spectral properties of the specific sibilant and do not represent the strength of unpleasant sibilants very well. Therefore the detected sounds are analyzed in more detail within the second step by calculating sharpness. To further reduce the complexity simplifications in computing sharpness are presented.

Finally sibilancy is reduced by filtering the disturbing part of the signal using cascaded linear-phase highpass- and lowpass filters and subtracting the filtered part. An equation for calculating a weighting factor and usable filter concepts are submitted.

Calculating Sharpness

Sharpness *S* is a weighted first moment of the critical-band rate distribution of specific loudness *N'*. [Zwi90]

$$S = c \bullet \frac{\int_{z=0}^{24Bark} N'(z) \bullet g(z) \bullet dz}{N} \; acum$$ (1)

where

$$N = \int_{z=0}^{24Bark} N'(z)\, dz$$ (2)

N is called the *loudness* of a signal. Sharpness is calibrated so that a narrow-band noise centered at 1 kHz produce a sharpness of 1 acum.

Aures modified equation (1) and described four steps for calculating sharpness.[Aur84] First the critical-band intensity has to be computed. This measure represents the energy of a signal within a critical-band (equal to 1 Bark), which is a frequency mapping related to the human hearing-system. A transformation from the frequency scale to the critical-band rate scale is

$$Z = \left[13 \bullet \arctan\left(0.76 \bullet \frac{f}{kHz} \right) + 3.5 \bullet \arctan\left(\frac{f}{7.5kHz} \right)^2 \right] Bark$$ (3)

Aures suggested an algorithm based on a 1024-point-FFT at a sample-rate of 30.72 kHz leading to 48 values from 0 to 24 Bark with an interval of 0.5 Bark. Taking the transmission factor *a_0* (representing the transmission between freefield and our hearing system) into account and considering a level-depended influence of neighboured critical-band intensities by adding slope excitations, excitation *E(z)* can be calculated in a second step. The third step calculates the specific loudness

$$N'(z) = N'_o \bullet \left(\frac{1}{s(z)} \frac{E_{TQ}(z)}{E_0} \right)^{0.23} \bullet \left[\left(1 - s(z) + s(z) \bullet \frac{E(z)}{E_{TQ}(z)} \right)^{0.23} - 1 \right] \frac{sone}{Bark}$$ (**4**)

where

$s(z) = 10^{(0.22 - 0.005 \bullet \frac{c}{Bark})} - 1$ *threshold factor* representing just-noticeable differences

E_0 excitation corresponding to the reference intensity $I_0 = 10^{-12} \frac{W}{m^2}$

$E_{TQ}(z)$ excitation at threshold in quiet

N'_0 reference specific loudness, which is used to calibrate the equation, so that a 1-kHz tone with a level 40dB produce exactly 1sone as total loudness.

Aures' modified equation for calculating sharpness is

$$S = c \bullet \frac{\int_{z=0}^{24Bark} N'(z) \bullet g'(z)\, dz}{\ln\left(\frac{N/sone + 20}{20}\right) sone} \; acum$$ **(5)**

where

$$g'(z) = \exp\left(\frac{0.171 \bullet z}{Bark}\right)$$ **(6)**

and

$$c = 0.08$$ (7)

Simplifications in calculating sharpness of speech signals on a DSP

Calculating sharpness on a DSP using the algorithm suggested by Aures is very inefficient. Besides it is impossible to measure the absolute level of the speech signals in the assigned equipment. Therefore the following simplifications are used.

1. As suggested by Zwicker [Zwi90] the slope excitations can be neglected in a first approximation of the specific loudness for speech signals. This is because of speech signals have a broad spectrum and slope excitations lead to significantly different excitations only for narrow spectrums.

2. The exponent 0.23 in equation (4) is debatable in technical literature and sometimes higher values are suggested. Therefore the exponent 0.25 (doubled square root) is used because of easier computation.

3. The terms *1-s(z)* and *-1* in equation (4) are neglected because they only have an effect in the presence of weak signals which will not be processed by the DeEsser algorithm.

4. Even though Aures showed that his equation for calculating sharpness is more accurate especially for signals with different loudness, equation (1) is used in the presented algorithm. Otherwise Aures' formula for the weighting factor *g(z)* (6) is still applied.

5. Instead of a 1024-point-FFT, a 256-point-FFT is used for calculating the critical-band intensities at 44.1kHz and 48 kHz sampling rates. At higher bands the smaller frequency resolution does not have any influence especially because of the broad spectrum of speech signals. Only a few lower bands cannot be computed because of missing spectral lines. For these bands a simple interpolation can be used. Additionally, the functions a_0 and *g(z)* are almost constant or linear in the lower frequency-region. Therefore detailed information about the distribution of energy within this region is not necessary.

Taking these simplifications into account calculating sharpness is reduced to the following steps:

1. 256-point-FFT
2. Computation of critical-band intensity *I(z)* for 48 overlapping critical bands by summing up the spectral lines and interpolating the missing values.
3. 48 multiplication's and 48 doubled square roots for calculating the specific loudness

$$N'(z) = \left[const. \bullet a_0(z) \bullet I(z)\right]^{0.25} \quad (8)$$

where a_0 can be calculated using the equation (derived from [Ter79])

$$\frac{a_0(f)}{db} = 6.5 \bullet e^{-0.6 \bullet (\frac{f}{kHz} - 3.3)^2} - 10 \, -3 \bullet \left(\frac{f}{kHz}\right)^4 \quad (9)$$

4. 48 multiplication's, 96 summation's and one division for calculating sharpness.

$$S = c \bullet \frac{\sum_{f=1}^{48} N'(\frac{1}{2}) \bullet g'(z)}{\sum_{f=1}^{48} N'(\frac{1}{2})} \quad acum \quad (10)$$

The mean error produced by this algorithm can be reduced by adjusting the factor c. Comparing the values calculated using this simplified algorithm with the values calculated by the original Aures-algorithm leads to a mean error of 0 % and a standard deviation of 5.8 % for 50 test sentences (in four different languages, spoken by native speakers). This accuracy is sufficient for detecting sibilants. A side effect of these simplifications is that the absolute level of the speech signals has not been taken into account in calculating sharpness.

Detecting Sibilants

Even though the simplified algorithm for calculating sharpness can be implemented on a DSP much more efficient than Aures' algorithm it cannot be computed for each sample on today's standard DSP-chips. On the contrary calculation of this value for non-overlapping blocks is possible in real-time but leads to an inaccurate detection of the beginning of a sibilant. Therefore two easily computable physical units can be utilized for a rough detection of sibilance:

1. zero-crossing rate *Z* and
2. short-time energy *E*.

Rabiner summarizes the coherence of speech signals with these units and specifies the following equations for computing.[Rab78]

$$Z_n = \sum_{m=-\infty}^{\infty} \left| \operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)] \right| w(n-m) \quad (11)$$

where

$$\begin{aligned} \operatorname{sgn}[x(n)] &= +1 \quad x(n) \geq 0 \\ &= -1 \quad x(n) \leq 0 \end{aligned} \quad (12)$$

and

$$\begin{aligned} \mathrm{w(n)} &= \frac{1}{2\mathrm{N}} \quad 0 \leq \mathrm{n} \leq \mathrm{N}-1 \\ &= 0 \qquad \text{otherwise} \end{aligned} \quad (13)$$

and

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)\, h(n-m) \quad (14)$$

where

$$\begin{aligned} e.g. \quad h(n) &= 0.54 - 0.46 \cos\left(\frac{2m}{N-1}\right) \; 0 \leq n \leq N-1 \\ &= 0 \qquad\qquad \textit{otherwise} \end{aligned} \quad (15)$$

He suggests a highpass filter starting at 60 Hz before calculating the zero-crossing rate (to minimize the error introduced by an offset) and a final lowpass filter (to flatten the output). A possible sibilant is detected if zero-crossing rate and short-time energy exceed certain thresholds. Observing the energy prevents false detection in case of silence whereas zero-crossing rates higher than 10-20% indicate the presence of sibilants.

If this algorithm notices a possible unpleasant sibilant, sharpness can be calculated leading to a more sophisticated two-step detector. (Figure 2) This detector does not increase the introduced time delay but decreases the basic computational load while improving the detection of a sibilant's beginning.

Reducing Sibilants

A basic algorithm for reducing sibilants is suggested in [Sap98]. The disturbing part of the signal is extracted using cascaded linear-phase highpass and lowpass filters, scaled by a weighting factor and subtracted from the correctly delayed input signal. The already computed critical band intensity represents aurally adequate information about the spectral properties of the sibilant. The frequencies nearest to the peak where the critical-band intensity falls below a certain percentage of the peak, can be taken as cut-off-frequencies for the highpass and lowpass filters. Percentages between 20% and 40% lead to good results, with smaller values increasing the bandwidth of the reduction. (Figure 3)

The choice of cut-off-frequencies should be restricted to the part of the critical-band intensity greater than 15 Bark. On the one hand this prevents instabilities in case of transients or voiced phonemes. On the other hand statistical evaluations on sibilants in four different languages showed that the low cut-off-frequencies will be in the region of 15 to 22 Bark and the high cut-off-frequencies will be in the region of 19 to 24 Bark so that the search-domain can be limited. (Figure 4)

The constraint of linear-phase filters reduces the number of usable filter implementations. Linear-phase IIR filters as introduced by Azizi in [Azi97a] and [Azi97b] as well as linear-phase FIR filters can be used, however considering signal delay introduced by the filtering process might be easier in case of linear-phase FIR filter implementations. Storing coefficients in memory is efficient because only 26 different coefficient-sets are necessary. The filters do not have to be adjusted during a sibilant as proven in [Sap98].

Filter parameters like order, stop-band attenuation, number of taps and steepness depend on the algorithm implementation and the hardware capabilities. Increasing stop-band attenuation (i.e., decreasing the influence on other signals than the disturbing part of the sibilant) increases the number of taps and the filter order as well. While flat flanks decrease the artifacts produced by subtracting the filtered part from the original signal, steep flanks minimize the influence to other signal components.

Bandwidth and center-frequency of the disturbing part of the sibilant as well as the calculated sharpness should be taken into account by computing a weighting-factor for the subtraction. One might use the following equation for calculating this factor:

$$W_{sub} = 1 - G \bullet (S \bullet A_S + B_\Delta \bullet A_B + (B_M - B_Z)A_M) \quad \textbf{(16)}$$

where

A_S, A_M, *AD* :weighting the influence of sharpness, center-frequency and bandwidth

S :sharpness

B_M :mid-frequency of cascaded filters

B_Z :'reference mid-frequency'

BD :bandwidth of cascaded filters

G :weighting the strength of de-essing

A time varying weighting-factor can be used for smoothly fading the DeEsser-algorithm in and out. Zero-crossing rate and short-time energy can be successfully utilized as criteria for the end of a sibilant by detecting when one of these units falls short of the threshold.

There are a lot of possibilities to provide experts with meaningful parameters. The thresholds of sharpness, zero-crossing rate and energy for detecting sibilants can be combined to create one threshold for the DeEsser. The percentages for finding the characteristic frequencies can be utilized to adjust the bandwidth of the filter as already mentioned and the parameter G in equation (16) is similar to the well known 'ratio knob'.

Conclusions

Major simplifications in calculating sharpness of speech signals were presented. Connected with zero-crossing rate and short-time energy this led to a two-step detector for sibilants which can easily be implemented on today's low-cost standard digital signal processors and which can be combined with an adaptive algorithm for reducing sibilants. These algorithms can be configured as a standalone unit without any parameter as well as a more sophisticated system including the well known parameters threshold, bandwidth and ratio. Several specialists like recording engineers and developers of professional audio equipment performed tests with these algorithms and suggested refinements of parameters.

REFERENCES

[Aur84] Wilhelm Aures: *Berechnungsverfahren für den Wohlklang beliebiger Schallsignale, ein Beitrag zur Gehörbez Schallanalyse*, Dissertation am Lehrstuhl für Elektroakustik der Technischen Universität München, 1984

[Azi97a] S.A. Azizi: *Realization of Linear Phase Sound Processing Filters Using Zero Phase IIR Filters*, Proc. 102nd AES Convention, Munich, Preprint 4506, March 1997

[Azi97b] S.A. Azizi: *Performance Analysis of Linear Phase Audio Filters based on the Zero Phase Filtering Concept*, Pro 103rd AES Convention, New York, Preprint 4535, September 1997

[Rab78] Lawrence R. Rabiner, Ronald W. Schafer: *Digital Processing of speech signals*, Prentice-Hall, New Jersey, 19

[Sap98] Markus Sapp, Martin Wolters, Jörg Becker-Schweitzer: *Reducing sibilants in recorded speech using psychoa models*, Paper accepted for presentation on ICA/ASA-Meeting, Seattle, 1998

[Ter79] E. Terhardt: *Calculating virtual pitch*, Hearing Research, Vol. 1, 155-182, 1979

[Wol97] Martin Wolters: *Entwicklung von Algorithmen zur Detektion und Unterdrückung von Zischlauten*, diploma t the RWTH Aachen , 1997

[Zwi90] E. Zwicker: *Psychoacoustics*, Springer-Verlag, Berlin, 1990

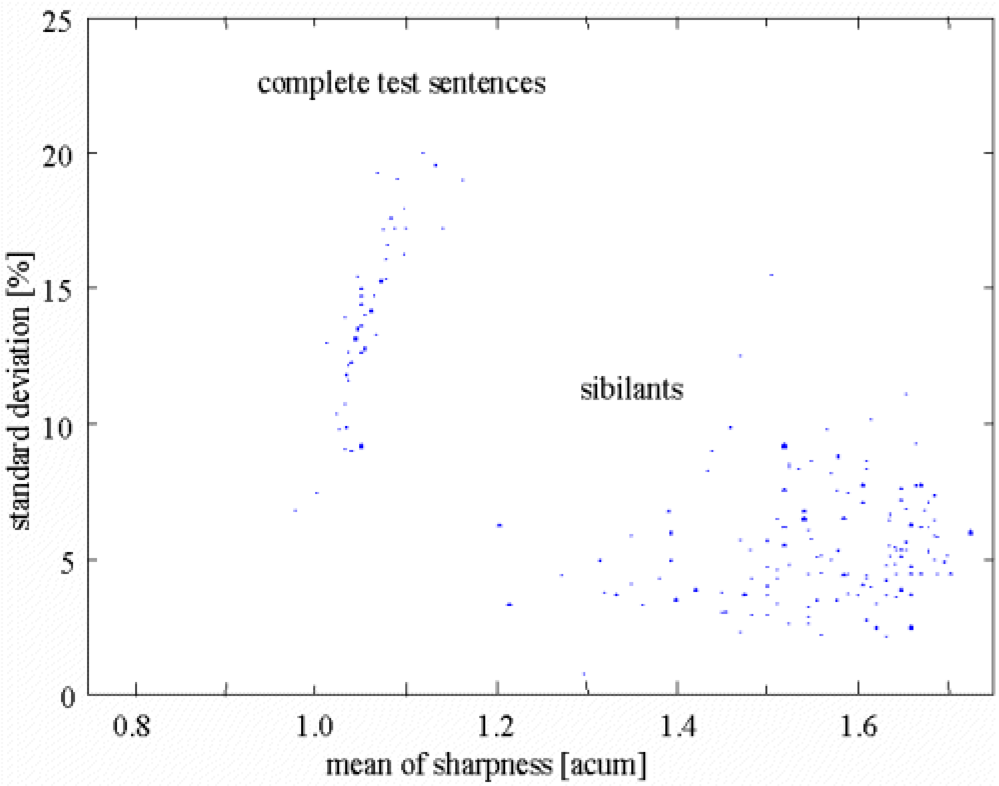

Figure 1. Correlation between sibilance and sharpness. 50 test sentences in German, English, Spanish and French were spoken by native speakers. The sibilants were marked as disturbing by at least three of four experts in a listening test. [Sap98]

Figures

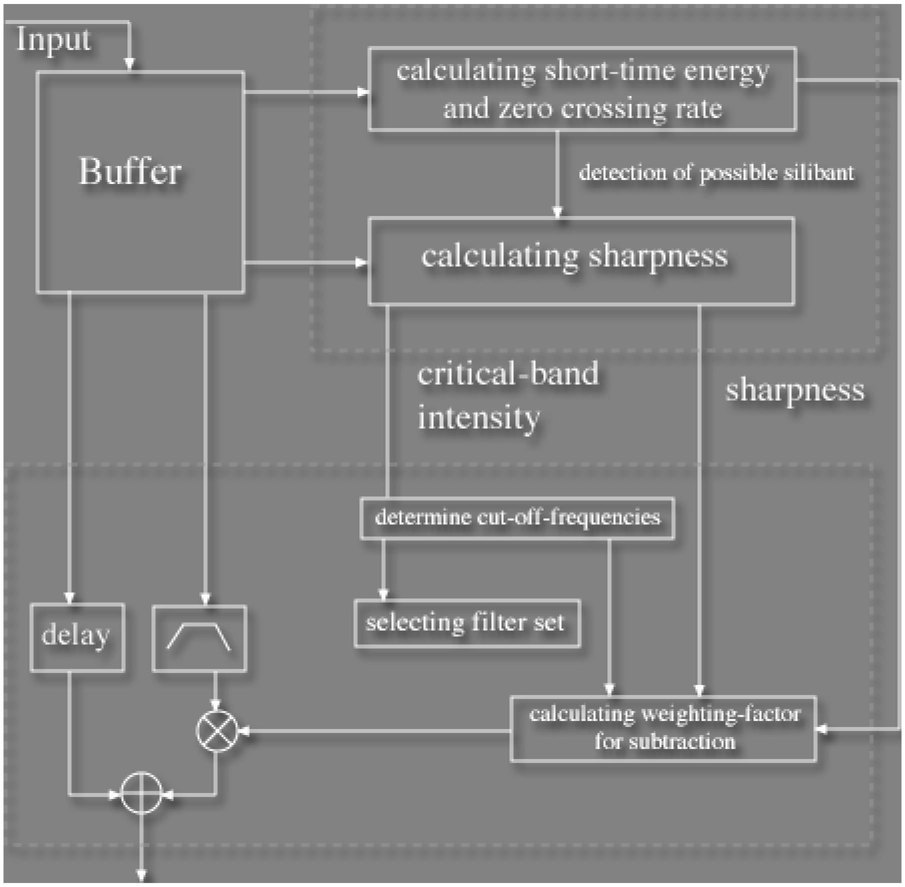
Input
Buffer
calculating short-time energy and zero crossing rate
detection of possible silibant
calculating sharpness
critical-band intensity
sharpness
determine cut-off-frequencies
selecting filter set
delay
calculating weighting-factor for subtraction

Figure 2. Two-step detector and algorithm for reducing sibilants.

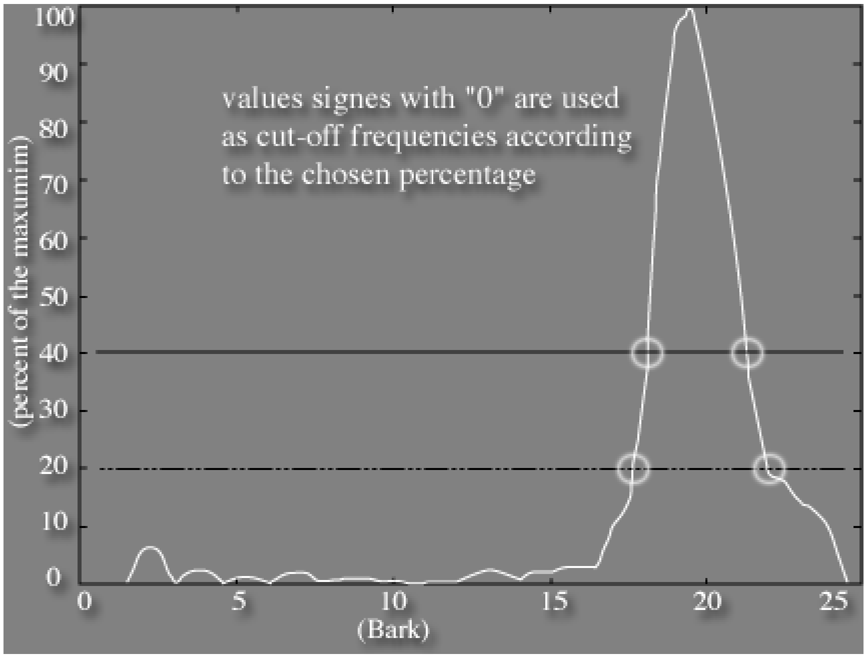

Figure 3. Typical example for using critical-band intensity to determine cut-off frequencies. Percentages between 20% and 40% lead to good results, with smaller values increasing the bandwidth of the reduction.

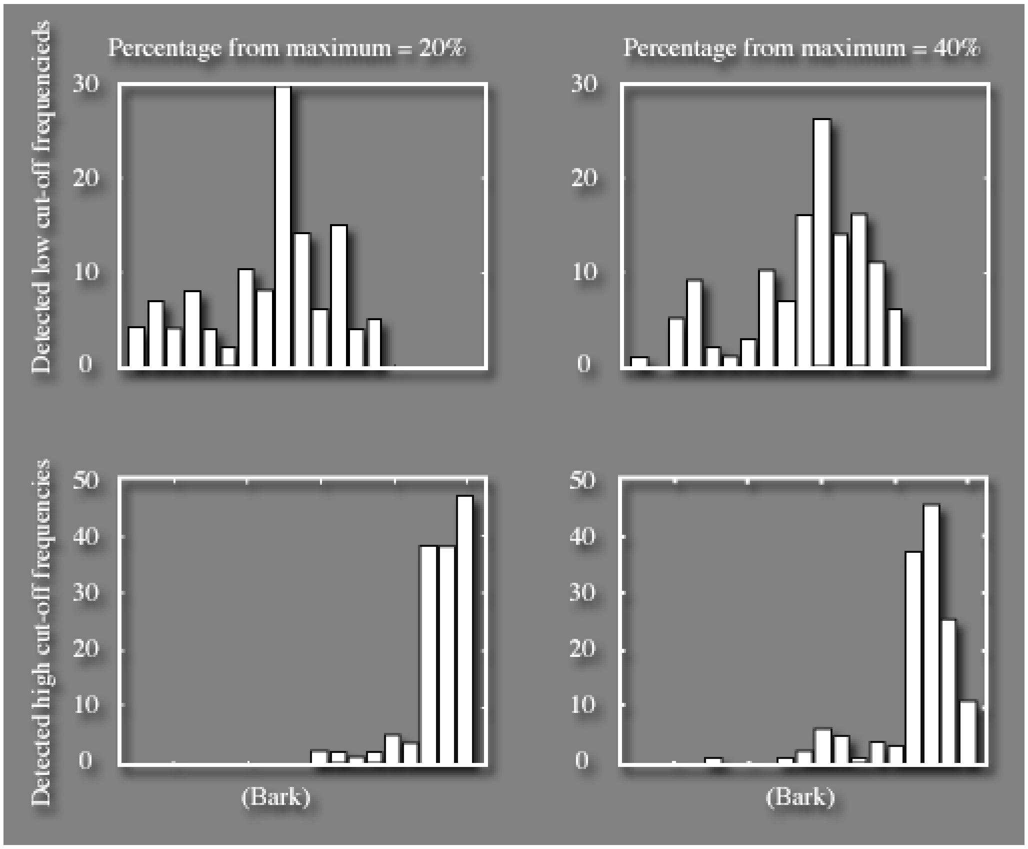

Figure 4. Detected cut-off frequencies for 141 sibilants in four different languages, spoken by male and female native speakers.