spgrambw: Plot Spectrograms in MATLAB

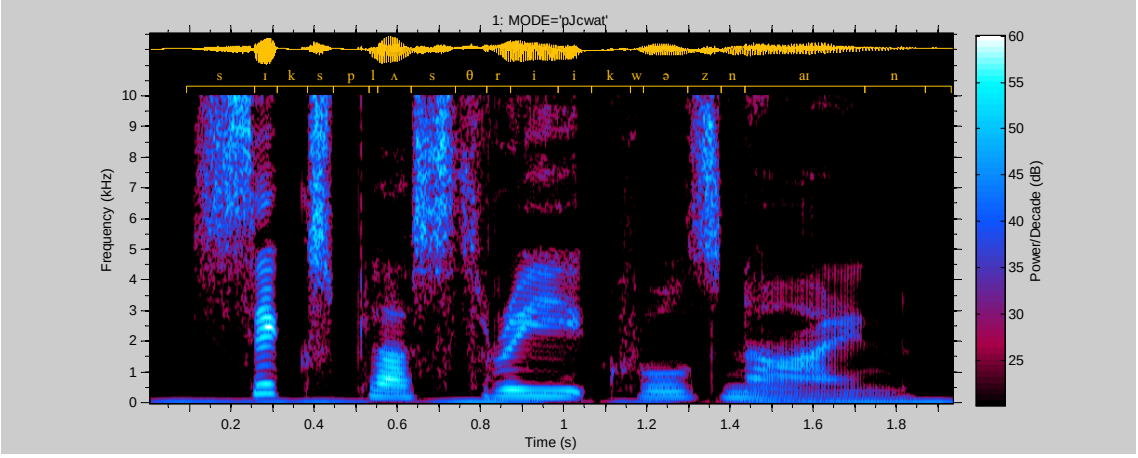
Mike Brookes

Version: 1639, 15/03/2012

Contents

1 Introduction

This document describes the spgrambw function which is part of the voicebox toolbox available at `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html` [Bro11]. We will use as an example, the following sentence "Six plus three equals nine" for which a spectrogram is shown below inculding the speech waveform and a time-aligned phonetic annotation.



2 Function call

The basic call to the function is:

[T,F,B]=spgrambw(S,FS,MODE,BW,FMAX,DB,TINC,ANN)

where all but the first two input arguments are optional. The input arguments are:

S input speech waveform

FS sample rate of speech waveform

MODE text string specifying a large range of options

BW the bandwidth of the spectrogram. This argument determines the tradeoff between time and frequency resolution.

FMAX specifies the range and resolution of the frequency axis

DB specifies the range of power spectral density displayed

TINC specifies the range and resolution of the time axis

ANN gives an optional annotation file containing words or phonemes.

If all you want to do is draw a spectrogram, then the function should be called without any output arguments. If output arguments are specified, then no spectrogram will be drawn unless the 'g' mode option is also given. The output arguments are

T gives the time of each time-axis sample point

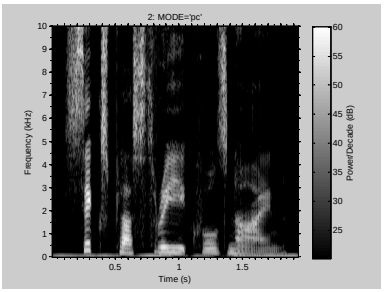
F gives the frequency of each frequency-axis sample point

B a 2-dimensional array giving the spectral density at each time-frequency point.

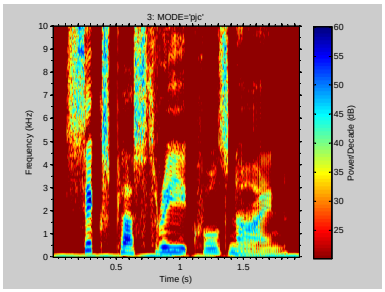
In the plots shown in this document, the title (above the spectrogram) shows the figure number (written {n} in the text), the value of the MODE argument and the value of any other arguments that are not null.

3 Colour maps

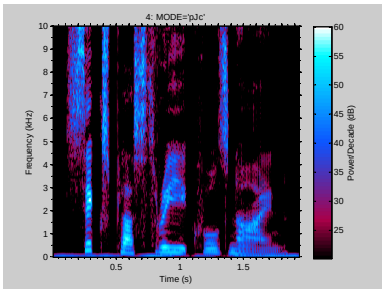
The default output is a monochrome spectrogram shown as {2}. Specifying the 'j' mode option uses the "jet" colourmap instead which is colourful and intuitive {3}. However it does not reproduce accurately if viewed or printed in monochrome and so I normally use the 'J' option instead which is less aggressive and converts accurately to monochrome {4}. Notice that I have also used the 'c' option in each case in order to include a colourbar giving the intensity scale in decibels.



2: Monochrome

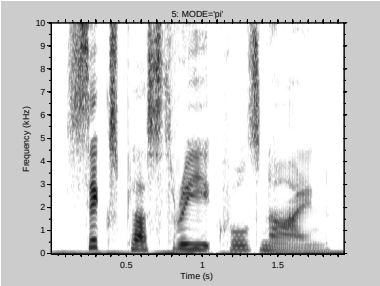


3: 'j'=Jet

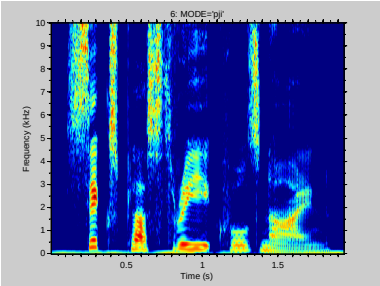


4: 'J'=Thermal

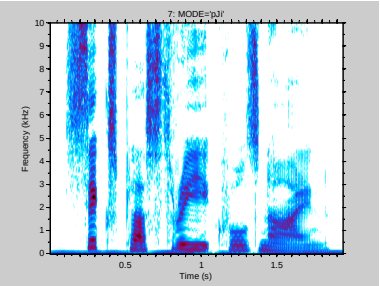
Adding the 'i' option inverts the colour map so that dark areas now correspond to high intensity. For these examples, I have omitted the 'c' option so the colourbar is missing.



5: 'i'= Inverted Monochrome



6: 'ij'=Inverted Jet

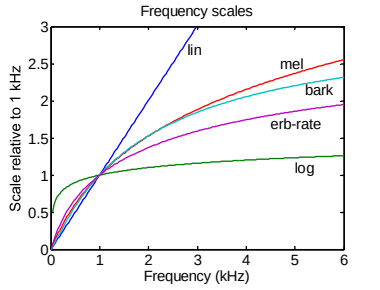


7: 'iJ'=Inverted Thermal

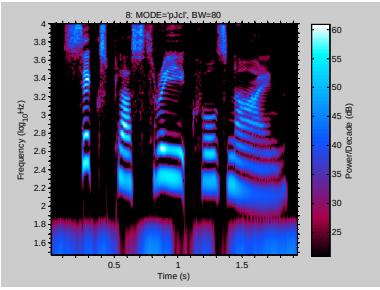
4 Frequency axis

4.1 Nonlinear frequency scaling

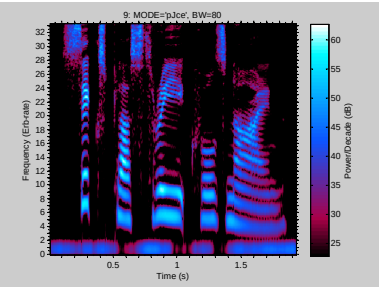
The default frequency axis is linear in Hz as seen in the examples above. Speech scientists usually prefer a nonlinear frequency scale in which high frequencies are compressed. There are several widely used frequency scales and these are plotted below (scaled to coincide at 1 kHz) [MG83, Ghi94, SVN37, Zwi61, ZT80]. The log scale {8} provides the most compression at high frequencies but it is more usual to use one of the physiological or psychoacoustical scales: Erb-rate {9}, Mel {10} or Bark {11}. The scale is selected by the MODE options 'l', 'e', 'm' or 'b'. In all cases, it is possible to add also the 'f' option which causes the frequency axis labels to be written in Hz as in {12}. In all the plots below, I have reduced the bandwidth to 80 Hz (see section 5) to give better frequency resolution.



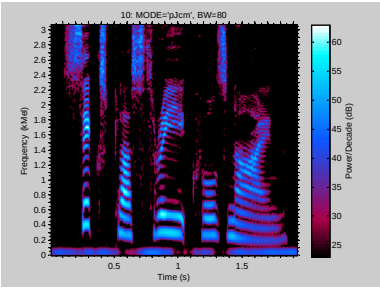
Frequency scales



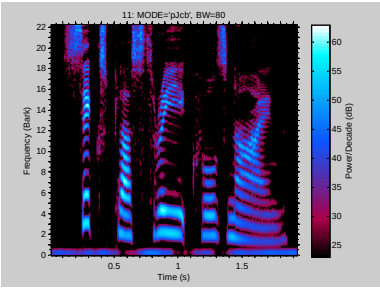
8: 'l' = Log scaled



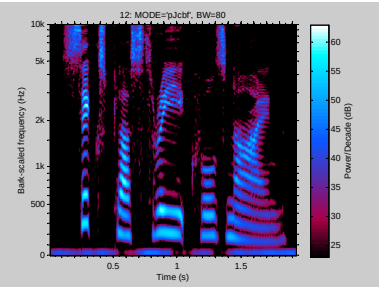
9: 'e' = Erb-rate scaled



10: 'm' = Mel scaled



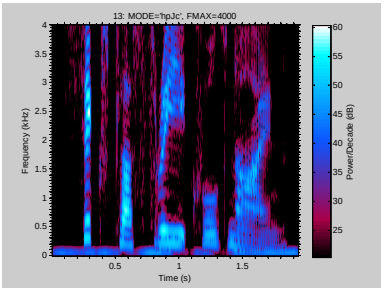
11: 'b' = Bark scaled



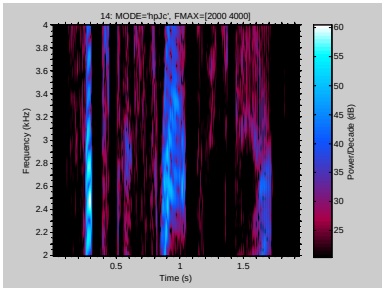
12: 'bf' = Bark + Hz labels

4.2 Frequency range and stepsize

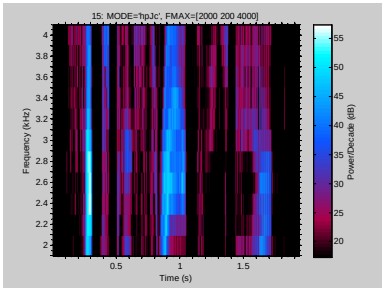
By default the frequency axis encompasses the entire range from 0 Hz to the Nyquist frequency, $\frac{1}{2}f_s$, but this is often too large. The FMAX input parameter allows you to specify the desired frequency range. Setting FMAX=4000 {13} restricts the frequency range to a maximum of 4 kHz while FMAX=[2000 4000] sets the range to 2 kHz to 4 kHz {14}. Normally the frequency stepsize is $\frac{1}{256}$ of the displayed range, but you can also specify the stepsize explicitly: FMAX=[2000 200 4000] goes from 2 kHz to 4 kHz in steps of 200 Hz {15}. If a nonlinear frequency scaling has been selected by the 'l', 'e', 'm' or 'b' options, then FMAX must be specified in scaled units unless the 'h' option is given, in which case they are in Hz as normal. Note that selecting a very small step size does not make the spectrogram any less blurry; the frequency resoulution is determined by the analysis bandwidth, BW, described in section 5.



13: 0 to 4 kHz



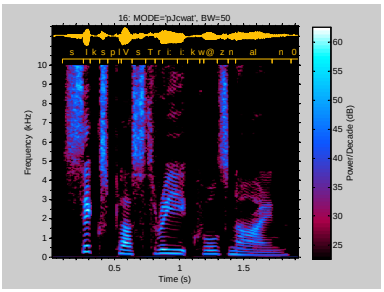
14: 2 to 4 kHz



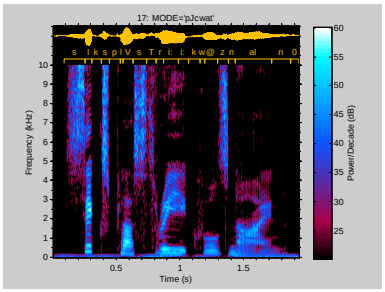
15: 200 Hz resolution

5 Analysis bandwidth

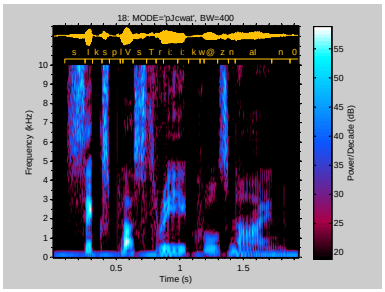
There is an unavoidable tradeoff between time resolution and frequency resolution that is often known as the "uncertainty principle". The BW input parameter specifies the -6 dB analysis bandwidth which is the frequency separation at which two tones will definitely give distinct peaks. From the point of view of frequency resolution, it follows that the smaller BW the better. However selecting a small value of BW means that rapid amplitude variations within any single frequency bin will be attenuated and, in particular, amplitude variations faster than $\frac{1}{2}$BW will be attenuated by more than -6 dB resulting in poor time resolution.



16: BW=50 Hz



17: BW=200 Hz (default)

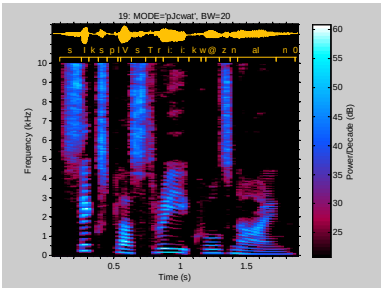


18: BW=400 Hz

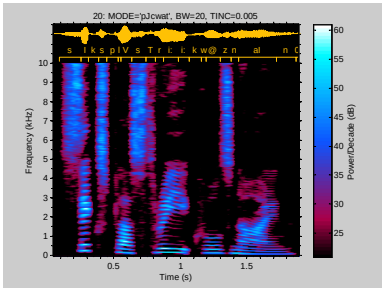
In this speech example, which is by a female talker, the larynx frequency varies from 300 Hz down to 150 Hz. If BW is chosen to be below the fundamental frequency, e.g. BW=50 Hz in {16}, the harmonics of the larynx frequency are clearly visible as quasi-horizontal stripes, however the time resolution is relatively poor. In a broadband spectrogram, in contrast, the bandwidth is chosen to be higher than the larynx frequency, e.g. BW=400 Hz in {18}, and the individual harmonics are no longer resolved. The time resolution is however much improved and it is possible to resolve the individual acoustic excitations arising from each larynx pulse; these are visible as vertical striations during the /aI/ phoneme of "nine" at a time of around 1.5 seconds. The default bandwidth is BW=200 Hz {17} which is often too large to reslve the larynx frequency harmonics but which makes the vocal tract resonances, or formants, easy to see.

6 Time Axis

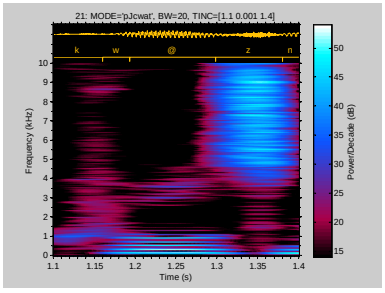
As discussed in section 5, the time resolution is determined by the BW parameter, and modulation frequencies above $\frac{1}{2}$BW are not shown in the spectrogram. For this reason, the default time-step is taken as $\frac{0.45}{\text{BW}}$ and, for small values of BW, this may give a blocky appearance {19}. To avoid this you can explicitly set a smaller time step using the TINC parameter as shown in {20}; note that although this results in a smoother appearance, it does not improve the time resolution which is still determined by the BW parameter (see section 5).



19: BW=20 Hz



20: BW=20, TINC=0.005



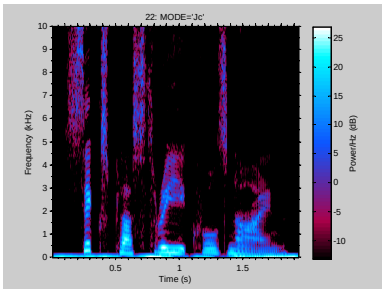
21: TINC=[1.1 0.001 1.4]

You can restrict the display to a specific time interval by setting TINC = $[t_{min}\, t_{max}]$ or TINC = $[t_{min}\, t_{step}\, t_{max}]$ if you want to specifiy the time-step as well {21}. Notice in {21} that the waveform and annotations remain correctly aligned.

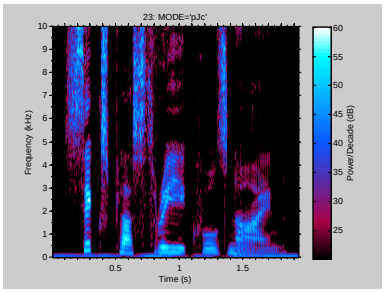
The sample time of $S(1)$ is assumed by default to be $T1 = \frac{1}{FS}$, but you can set it to any other value by making the second input argument a vector: [FS T1].

7 Intensity scaling

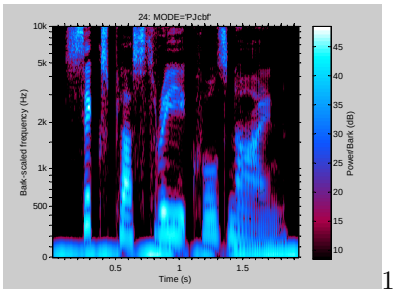
The default spectrogram shows the spectral density in units of "power per Hz" {22}. Because most speech energy is concentrated at low frequencies, this can make it difficult to see detail in the display at both low and high frequencies. To avoid this, you can use the 'p' option to display "power per decade" instead: this option multiplies the power by a value proportional to the frequency and so emphasises high frequencies {23}. If you are using one of the non-linear frequency scaling options described in section 4.1, you have a third option which is to show "power per bark/erb/..." {24}.



22: Power/Hz

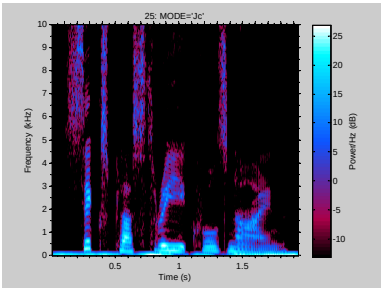


23: 'p'=Power/Decade

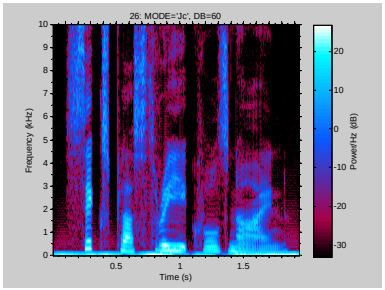


24: 'P'=Power per Bark

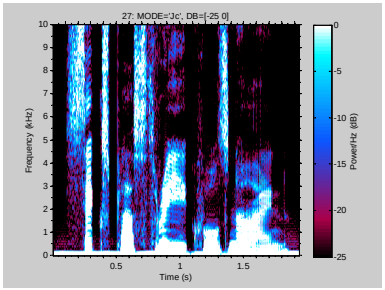
Normally, the display shows a range of 40 dB from the maximum power anywhere in the spectrrogram {25}. You can change this to a different range by setting the DB parameter either to the desired range{26} or alternatively to the minimum and maximum powers to display: DB = $[P_{min}\, P_{max}]$ {27}. This option is especially useful if you want to have several spectriograms with identical displayed power ranges. Values outside the selected range will be set to either the minimum or maximum.



25: 40 dB range (default)



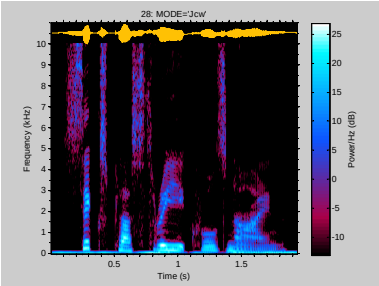
26: DB=60



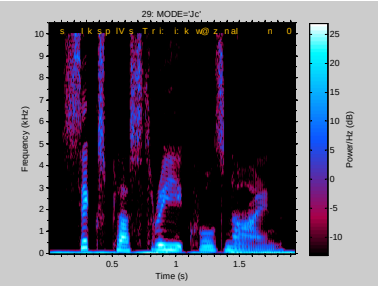
27: DB=[-25 0]

8 Waveform and transcription

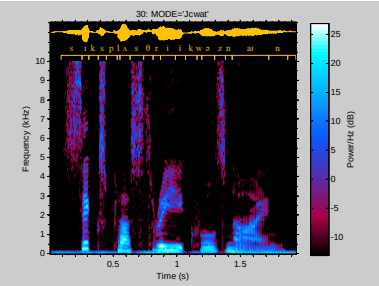
It is often helpful to display the time-domain waveform on the spectrogram and you can do so with th 'w' option {25}. If you have a transcription or other time-aligned annotation, you can specify it as the ANN input. Each row of the ANN cell array is of the form {$[t_{start}\, t_{end}]$ 'text'}. By default, the annotations are left-aligned within their time intervals without any time markers {26}. If you want to display phonetic characters, you will need to install a non-unicode IPA font such as the SIL93 fonts (available for download from the Voicebox website). You can specify the font of each annotation entry by including a third column; each row of ANN is now of the form {$[t_{start}\, t_{end}]$ 'text' 'font'}. Example {27} uses the 'SILDoulos IPA93' font and also includes the options 'a' which centres the annotations in their time interval and 't' which includes time markers.



25: 'w'=show waveform



26: ANN input



27: 'wat' + ANN font

9 Output Arguments

Specifying output arguments normally suppresses the spectrogram plot unless the 'g' option is given. Note that, perhaps unexpectedly, the spectrogram array is the third output rather than the first.

If you save the B output (with a linear frequency scale and without the 'p' or 'P' options), you can use it as the input to a subsequent call to spgrambw instead of a time-domain waveform. In this case FS=[FS T1 FINC F1] where FS is now the frame rate (each frame is one row of B), T1 is the time of the first row of B, FINC is the frequency increment and F1 is the frequency of the first column in B.

10 MODE string options

a centre-align annotations rather than left-aligning them

b bark scale

c include a colourbar as an intensity scale

d give the **B** ouput in decibels rather than in power.

D clip the output B array to the limits specified by the "db" input

e erb scale]

f label frequency axis in Hz rather than mel/bark/...

g draw a graph even if output arguments are present

h units of the FMAX input are in Hz instead of mel/bark/... In this case, the Fstep parameter is used only to determine the number of filters.

H express the F output in Hz instead of mel/bark/...

i inverted colourmap" (white background)

j jet colourmap

J "thermal" colourmap that is linear in grayscale. Based on Oliver Woodford's % real2rgb at http://www.mathworks.com/mat

l log10 Hz frequency scale

m mel scale

p calculate "power per decade" rather than "power per Hz". This effectively increases the power level at high frequencies and so maes them more visible

P calculate "power per erb/mel/..." rather than power per Hz.

t add time markers with annotations

w draw the speech waveform above the spectrogram

11 MATLAB Code for figures

The following code was used to generate all the figures in this document:

```
% demonstrations for the spgrambw tutorial
clear all;
close all;
p={1 2.5 'pJcwat' [] [] [] [] 2;
    2 1.33 'pc' [] [] [] [] 0;
    3 1.33 'pjc' [] [] [] [] 0;
    4 1.33 'pJc' [] [] [] [] 0;
    5 1.33 'pi' [] [] [] [] 0;
    6 1.33 'pji' [] [] [] [] 0;
    7 1.33 'pJi' [] [] [] [] 0;
    8 1.33 'pJcl' 80 [] [] [] 0;
    9 1.33 'pJce' 80 [] [] [] 0;
    10 1.33 'pJcm' 80 [] [] [] 0;
    11 1.33 'pJcb' 80 [] [] [] 0;
    12 1.33 'pJcbf' 80 [] [] [] 0;
    13 1.33 'hpJc' [] [4000] [] [] 0;
    14 1.33 'hpJc' [] [2000 4000] [] [] 0;
    15 1.33 'hpJc' [] [2000 200 4000] [] [] 0;
    16 1.33 'pJcwat' [50] [] [] [] 1;
    17 1.33 'pJcwat' [] [] [] [] 1;
    18 1.33 'pJcwat' [400] [] [] [] 1;
    19 1.33 'pJcwat' [20] [] [] [] 1;
    20 1.33 'pJcwat' [20] [] [] [0.005] 1;
    21 1.33 'pJcwat' [20] [] [] [1.1 0.001 1.4] 1;
    22 1.33 'Jc' [] [] [] [] 0;
    23 1.33 'pJc' [] [] [] [] 0;
    24 1.33 'PJcbf' [] [] [] [] 0;
    25 1.33 'Jc' [] [] [] [] 0;
    26 1.33 'Jc' [] [] [60] [] 0;
    27 1.33 'Jc' [] [] [-25 0] [] 0;
    28 1.33 'Jcw' [] [] [] [] 0;
    29 1.33 'Jc' [] [] [] [] 1;
    30 1.33 'Jcwat' [] [] [] [] 2};
yfig=420; % height of figures
emf=1;      % set to 1 to print
args={'BW' 'FMAX' 'DB' 'TINC'};

% read the SFS-format speech file

fn='at05f0.sfs';
[sp,fs]=readsfs(fn,1,1); % speech signal
[pt,fw]=readsfs(fn,5,2);    % phonetic transcription
ann=[mat2cell([cell2mat(pt(:,1)) cell2mat(pt(:,1:2))*[1;1]]/fw,ones(1,size(pt,1))) pt(:,3)
ipa=ann(:,[1 2 2]);
ipa(:,2)={'s' 'I' 'k' 's' 'p' 'l' 'Ã' 's' 'T' 'r' 'i' 'i' 'k' 'w' '■' 'z' 'n' 'aI' 'n' ' '
ipa(:,3)=repmat({'SILDoulos IPA93'},size(ipa,1),1);


for i=1:size(p,1)
    if p{i,1}>0
        figure(p{i,1})
        set(gcf,'Position',[100 100 round(yfig*p{i,2}) yfig],'InvertHardcopy','off');
        switch p{i,8}
            case 0
                spgrambw(sp,fs,p{i,3},p{i,4},p{i,5},p{i,6},p{i,7});
            case 1
                spgrambw(sp,fs,p{i,3},p{i,4},p{i,5},p{i,6},p{i,7},ann);
```

```
            case 2
                spgrambw(sp,fs,p{i,3},p{i,4},p{i,5},p{i,6},p{i,7},ipa );
        end
        ss=sprintf('%d: MODE=''%s''',p{i,1},p{i,3});
        for j=4:7
            if numel(p{i,j})==1
                ss=sprintf('%s, %s=%g',ss,args{j-3},p{i,j});
            elseif numel(p{i,j})>1
                ss=sprintf('%s, %s=[%s',ss,args{j-3},sprintf('%g ',p{i,j}));
                ss=[ss(1:end-1) ']'];
            end
        end
        title(ss);
        if emf, eval(sprintf('print -dmeta %s',sprintf('%s%d',mfilename,round(gcf)))); end
        if i>1 && i<28
            close(i);
        end
    end
end

% now plot other graphs

for i=201:201
    figure(i)
    switch i
        case 201
            fax=linspace(0,6000,200)';
            y=[fax [nan; log10(fax(2:end))] frq2mel(fax) frq2bark(fax) frq2erb(fax)];
            [v,iv]=min(abs(fax-1000));
            y=y./repmat(y(iv,:),length(fax),1);
            plot(fax/1000,y);
            set(gca,'ylim',[0 3]);
            xlabel('Frequency (kHz)');
            ylabel('Scale relative to 1 kHz');
            title('Frequency scales')
            txt={2.8 2.7 'lin'; 5 1.1 'log'; 4.7 2.5 'mel'; 5.2 2.15 'bark'; 4.5 1.7 'erb-
            for j=1:5
                text(txt{j,1},txt{j,2},txt{j,3})
            end
            figbolden
    end
    if emf, eval(sprintf('print -dmeta %s',sprintf('%s%d',mfilename,round(gcf)))); end
    close(i);
end
```

References

[Bro11] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," Imperial College, Software Library, 2011. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[Ghi94] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans Speech Audio Processing*, vol. 2, no. 1, pp. 115–132, 1994.

[MG83] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Amer.*, vol. 74, pp. 750–753, 1983.

[SVN37] S. S. Stevens, J. Volkman, and E. B. Newman, "A scale for the measurement of the psychological magnitude of pitch," *J. Acoust. Soc. Amer.*, vol. 8, pp. 185–19, 1937.

[ZT80] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Amer.*, vol. 68, no. 5, pp. 1523–1525, Nov. 1980.

[Zwi61] E. Zwicker, "Subdivision of audible frequency range into critical bands," *J. Acoust. Soc. Amer.*, vol. 33, p. 248, 1961.