Intelligibility of Speech

2ND SEMESTER PROJECT
DTU ELECTRICAL ENGINEERING
ACOUSTIC TECHNOLOGY
SPRING SEMESTER 2008

Group 5
Troels Schmidt Lindgreen – s073081
David Pelegrin Garcia – s070610
Eleftheria Georganti – s073203

Instructor
Thomas Ulrich Christiansen
Title:
Intelligibility of Speech

Course:
31236 Auditory Signal Processing and Perception
Spring semester 2008

Project group:
5

Participants:
Troels Schmidt Lindgreen
David Pelegrin Garcia
Eleftheria Georganti

Supervisor:
Torsten Dau

Instructor:
Thomas Ulrich Christiansen

Date of exercise: April 17th

Pages: 27

Copies: 4

Synopsis:
This report deals with the intelligibility of speech and objective methods that can be used to predict it. Three different methods are used to estimate speech intelligibility. Namely, Articulation Index (AI), Speech Intelligibility Index (SII) and Speech Transmission Index (STI).

It has been concluded that it is possible to determine the gender of the speaker, but not the spoken language from a long-term magnitude spectrum. The main peak in the spectrum corresponds to the fundamental frequency, which is around 125 Hz for male speakers and 250 Hz for females. AI and SII are useful estimators of speech intelligibility for band-limited systems whereas STI can only be calculated if the impulse response is available and it is used mainly in room acoustics measurements.

No part of this report may be published in any form without the consent of the writers.
Introduction

This exercise investigates the intelligibility of speech and three different methods to estimate it are examined. Namely, the Articulation Index (AI), the Speech Intelligibility Index (SII) and the Speech Transmission Index (STI).

The magnitude spectrum of speech and speech shaped noise of four existing speech samples is examined. It is investigated whether regional and gender characteristics can be obtained through these spectra.

A MATLAB function is composed in order to calculate the Articulation Index (AI) and the Speech Intelligibility Index (SII). Low-pass and high-pass transmission channels with different cut-off frequencies are used. The predictions are plotted as a function of the cut-off frequency. Further investigation of the influence of the speech level on AI and SII calculation is carried out.

In the last part, the presence of non-linear distortions in the transmission channel is investigated and the concept of predicting the speech intelligibility with the Speech Transmission Index (STI) is used. Simulated impulse responses are used and the corresponding STI is calculated.
Chapter 1

Theory

To get a better understanding in speech generation and perception, methods have been developed in order to determine the speech intelligibility threshold. Speech intelligibility indicates how many correct words the listener understands from speech in a given situation. Speech is generated from a talker and is received by a listener, as illustrated in figure 1.1.

Figure 1.1: Speech from one person to another. Inspired from [Poulsen, 2007, p. 3]
1.1 Definition of speech intelligibility

Speech intelligibility is a number that indicates how much speech is understood correctly in a certain situation. Speech intelligibility can be formulated as:

\[ \text{Speech intelligibility} = \frac{100}{T} \cdot R \]  \hspace{1cm} (1.1)

Where:
- \( T \) is the number of speech units in the test.
- \( R \) is the number of correct speech units.

Speech intelligibility is defined as the percentage of speech units that are understood correctly from a speech intelligibility test [Poulsen, 2005, p. 62]. These units can be words (and equation (1.1) is referred to as word score), syllables, or other units. Speech intelligibility is typically drawn as a function of signal-to-noise ratio, as shown in figure 1.2.

1.2 Speech reception threshold

The Speech Reception Threshold (SRT) is defined as the signal-to-noise ratio at which the psychometric function that describes the speech intelligibility is equal to 0.5. Thus, the SRT depends on the test subject, the material and the score units used to determine the intelligibility.

![Figure 1.2: Word score from Dantale as a function of signal-to-noise ratio. Redrawn from [Poulsen, 2005, p. 67].](image)

A typical speech intelligibility test consists of a number of words being read out aloud by a speaker. The listener writes what he/she hears, and the lists are compared afterwards.
to find the speech intelligibility, see figure 1.3

Figure 1.3: Elements involved in speech intelligibility test. Inspired by [Poulsen, 2005].

There are a lot of factors to consider in speech communication systems. The transmission system is an important part of speech intelligibility and it can be anything from the room which is spoken in, a telephone line or the ears of the listener. It depends on the situation, where the transmission system can be influenced by reverberation, distortion or noise etc.

Figure 1.3 shows that the transmission system is only a small part of the system. Other important factors are:

- **Word material**: Could be sentences, words or numbers.
- **Presentation method**: Single words presented or a small text/sentence.
- **Open/closed response test**: Whether or not the words are known to the listener beforehand.
- **Speaker**: Speed of talking, pronunciation and dialect. Normal speech, raised voice or shouting.
- **Listener**: Hearing ability, training, vocabulary, dialect, mother tongue.
- **Scoring method**: Scoring by exact word or phoneme. Oral or written answers.

### 1.3 Speech shaped noise

Speech Shaped Noise (SSN) is defined as a random noise that has the same long-term spectrum as a given speech signal. There are many techniques to produce this kind of noise, e.g.:

- Gaussian noise can be equalized to have the same magnitude spectrum as a given speech signal.
• The phase of a speech signal can be randomized. In this way, the signal will be random noise but it will keep the magnitude spectrum of the original speech.

• Multiple replicas of the signal (reversed or not) can be delayed and added together. However, these techniques do not preserve the intrinsic fluctuations of speech and thus the modulations are reduced. Nevertheless, it is possible to produce speech shaped noise that keeps the envelope spectrum of the speech. According to [Dau and Christiansen, 2008, p. 4], one example of this kind of shaped noise is found in the so-called ICRA-noises, which are created by the following stages, illustrated in figure 1.4:

1. Filtering of the speech signal in three adjacent bands covering the whole frequency range of the signal.

2. Randomly reversing the sign of the samples in each band. This process modifies the spectrum but keeps the envelope.

3. Filtering of each band with the same filters as in stage 1.

4. Add and normalise the contributions from the three bands.

Figure 1.4: Block diagram of the ICRA-noise generation for a given speech signal \( s(t) \). \( r_1(t) \), \( r_2(t) \) and \( r_3(t) \) are uncorrelated random signals that take values \(-1\) or \(1\) and have the same length as the original signal.

### 1.4 Articulation index

The Articulation Index (AI) was developed in order to evaluate how noise over a telephone line affects speech intelligibility. AI estimates the physical signal-to-noise ratio as a correlated internal representation in the listener by calculating an “effective signal-to-noise ratio” for 20 frequency bands. Masking effects are taken into account in the calculation such that the noise may mask adjacent frequency bands. The effective contributions of each frequency band are weighted using a band importance function to yield an index value from 0 to 1, where 1 marks maximum intelligibility and 0 marks no intelligibility. [Dau
It must be noted that AI quantifies the degradations of the signal introduced by the transmission channel only. In order to predict speech intelligibility, knowledge about the specific task is required. It is then possible to read off the speech intelligibility from AI graphs which are based on speech intelligibility measurements representing average population responses of normal hearing talkers and listeners in stationary background noise. Such a graph can be seen on figure 1.5.

![Figure 1.5: Relation between speech intelligibility (in %) and AI. From [Poulsen, 2005, p. 76](#).](image)

AI is standardized in ANSI S3.5-1969.

### 1.5 Speech intelligibility index

The speech intelligibility index (SII) is based on the AI principle. SII uses other weighting function and a number of modifications are implemented. One of the modifications is the correction for change in speech spectrum according to vocal effort.

SII in standardized in ANSI S3.5-1997.

### 1.6 Speech transmission index

An approach to quantify the speech intelligibility from the objective sound field data is based on the Modulation Transfer Function (MTF). MTF is a function $m(\Omega)$ (where
\[ \Omega = i \cdot \omega \] which can be calculated from the impulse response \( h \) of the transmission system and equals to:

\[
m(\Omega) = \frac{\int_0^\infty h^2(t)e^{-i\omega t}dt}{\int_0^\infty h^2(t)dt}
\] (1.2)

In other words equation (1.2) is the complex Fourier Transform of the squared impulse response divided by its total energy. The MTF quantifies the leveling effect of reverberation on the envelope of speech signals. [Houtgast and Steeneken, 1973] have developed a procedure to convert MTF data measured in seven octave bands and at several modulation frequencies into one single quantity, which they called “Speech Transmission Index (STI)”. This conversion involves averaging over a certain range of modulation frequencies and takes into account the contribution of the various frequency bands to speech quality and the masking effect between adjacent frequency bands that occurs in the hearing system.

The exact calculations needed to determine the STI requires firstly, the calculation of the complex MTF for the octave bands centered at 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, and 8 kHz, according to equation (1.2). Then a reduction of modulation \( m_{F,k} \) for the different 14 modulation frequencies (0.63 Hz, 0.8 Hz, 1 Hz, 1.25 Hz, 1.6 Hz, 2 Hz, 2.5 Hz, 3.15 Hz, 4 Hz, 5 Hz, 6.3 Hz, 8 Hz, 10 Hz, and 12.5 Hz) is required, see figure 1.6. This equals to a transformation to a signal-to-noise ratio:

\[
\text{SNR}_{F,k} = 10 \log_{10} \left( \frac{m_{F,k}}{1 - m_{F,k}} \right)
\] (1.3)

where \( \text{SNR}_{F,k} \) is the signal-to-noise ratio in a given combination of an octave band \( F \) and a modulation frequency band \( k \) and \( m_{F,k} \) is the modulation frequency of the corresponding combination of bands.

The 14 values for each octave band are then averaged across modulation frequency yielding 7 values that correspond to an octave band. These values are then multiplied by a factor \( a_F \):

\[
\text{SNR}_F = a_F \sum_{k=1}^{14} \frac{\text{SNR}_{F,k}}{14}
\] (1.4)

Afterwards, the 7 octave band values \( \text{SNR}_F \) are limited to ±15 dB, averaged and then normalized. This leads to the expression of the STI:

\[
\text{STI} = \frac{\text{SNR}_{av} + 15}{30}
\] (1.5)

where \( \text{SNR}_{av} \) corresponds the average value of the SNR over the 7 octave bands. STI scores are rated as seen in table 1.1.
modulation depth will be reduced at the receiving point if noise or reverberation has been present in the course from sender to receiver. The reduction of the modulation is described by the MTF, \( m(F) \), which is a function of the modulation frequency, \( F \). MTF is the basis for the calculation of STI.

A noise signal with the same long term spectrum as speech (speech shaped noise) is used in the measurement of STI. The noise signal is divided into 7 octave bands from 125 Hz to 8 kHz, see figure 11.2. Each octave band is separately modulated with 14 modulation frequencies (one at a time) from 0.63 Hz to 12.5 Hz. The modulation is sinusoidal and 100\% and the signal is emitted from the sender (the speaker).

For each of the 98 combinations of octave band and modulation frequency the reduction in modulation is determined. This is the MTF as a function of octave band centre frequency, \( k \), and modulation frequency, \( F \). The reduction in modulation, \( m(F,k) \), is transformed to a signal-to-noise ratio by means of

\[
\text{SNR}_{F,k} = 10 \log \left( \frac{m}{1-m} \right) \text{ dB}
\]

As in the AI method the SNR is truncated to a dynamic range of 30 dB but here the limits are ±15 dB (in AI the limits are +12 dB and –18 dB). If SNR is >15 dB then SNR is set to 15 dB and if SNR is < –15 dB then SNR is set to –15 dB.

Like in AI, the relation between STI and speech intelligibility score is not linear. It is possible to read off the speech intelligibility from STI graphs like figure 1.7.

STI is standardized in IEC 286-16-1988, and has been revised in 2003.
and in the STI procedure. It is seen that especially around 2 kHz the weights are different. The STI in itself is not enough to predict the speech intelligibility. In order to use the STI it is necessary to know the relation between speech intelligibility and STI. This is illustrated in figure 11.4 for some typical word materials.

Figure 11-4. Relation between intelligibility and STI for different word materials. Compare with Figure 10-6.

11.2 RASTI

As seen in the STI section it is a somewhat complicated matter to calculate STI. This is not satisfying for an objective method that should replace the cumbersome subjective measuring methods. Therefore a simpler method, RASTI (= RApid STI) has been developed based on the same principles as in STI. In RASTI the number of combinations of octave bands and modulation frequencies are reduced from 98 to 9. In Figure 11-2 the combinations of frequency bands and modulation frequencies used in RASTI are shown in grey. Only the octave bands 500 Hz and 2 kHz are used and the modulation frequencies are selected so that they cover the most important range.

Figure 11.5 show the envelopes curves for the 500 Hz and the 2 kHz noise bands. Figure 11.6 shows the calculation procedure. There are no weighting factors in the RASTI calculation but because of the 5 modulation frequencies used in the 2 kHz band and only 4 modulation frequencies in the 500 Hz band, there will be a slightly higher weight for the 2 kHz band in the average calculation.

Figure 1.7: Relation between intelligibility and STI for different word materials. From [Houtgast and Steeneken, 1985, p. 11].
Chapter 2

Results

2.1 Spectrum of speech

Four speech recordings are listened and their full bandwidth long-term magnitude spectrum is calculated with the script provided in appendix A.

The long-term magnitude spectrum from 100 Hz to 10 kHz of the 4 speech signals and the corresponding smoothed spectrum (using the supplied script Oct3Smooth.m) can be seen in figures 2.1, 2.2, 2.3 and 2.4.

Figure 2.1: Long-term magnitude spectrum from 100 Hz to 10 kHz and the corresponding smoothed spectrum for a danish female speaker.

Figure 2.2: Long-term magnitude spectrum from 100 Hz to 10 kHz and the corresponding smoothed spectrum for a danish male speaker.
The sampling frequency \( f_s \), the number of samples \( N \), the bandwidth \( B \) and the frequency resolution \( F_{\text{res}} \) of the above speech signals are related when performing an FFT. More precisely the frequency resolution \( F_{\text{res}} \) and the bandwidth of the transformed signal \( B \) equal to:

\[
F_{\text{res}} = \frac{1}{N} \quad (2.1)
\]

\[
B = \frac{f_s}{2} \quad (2.2)
\]

### 2.2 Speech shaped noise

The script `randomizePhase` is used to generate speech shaped noise from the 4 speech samples. This script assigns a random value to the phase of each frequency component of the signal. As explained in section 1.3, this kind of procedure does not preserve the modulation of the signal. As an example, in figure 2.5 a speech signal and the speech shaped noise calculated with this method are shown, both in the time domain and the magnitude spectrum.
2.3 High-pass and low-pass filtered speech

In order to determine the importance of each frequency region for the intelligibility of speech, low-pass and high-pass transmission channels with different cut-off frequencies can be used.

High-pass and low-pass filters with 20 different cut-off frequencies in the frequency range between 100 Hz and 10 kHz were calculated using Matlab (see appendix A). The transfer functions of the high-pass filters with different cut-off frequencies have been plotted and can be seen in figure 2.6. Additionally, the transfer functions of the low-pass filters with different cut-off frequencies have been plotted and can be seen in figure 2.7.

For the low-pass and high-pass filters the AI and SII are calculated using the Matlab code that can be found in appendix A. The speech spectrum level is assumed to be “normal”, “raised”, “loud” or “shout”, corresponding to the SPL values shown in table 2.1. The results for the high-pass and low-pass filters can be seen in figures 2.8 and 2.9 respectively.
Figure 2.6: Transfer functions of high-pass filters with different cut-off frequencies in the frequency range between 100 Hz and 10 kHz.

<table>
<thead>
<tr>
<th>Voice level</th>
<th>dB SPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>63</td>
</tr>
<tr>
<td>raised</td>
<td>68</td>
</tr>
<tr>
<td>loud</td>
<td>75</td>
</tr>
<tr>
<td>shout</td>
<td>82</td>
</tr>
</tbody>
</table>

Table 2.1: Overall speech sound pressure levels for a female speaker [Poulsen, 2005].
Figure 2.7: Transfer functions of low-pass filters with different cut-off frequencies in the frequency range between 100 Hz and 10 kHz.

Figure 2.8: AI and SII for high-pass filters as a function of the cutoff frequency when the speech spectrum level is assumed to be (a) “normal”, (b) “raised”, (c) “loud” and (d) “shout”.

Page 13
Figure 2.9: AI and SII for low-pass filters as a function of the cutoff frequency when the speech spectrum level is assumed to be (a) “normal”, (b) “raised”, (c) “loud” and (d) “shout”.
2.4 Speech intelligibility in simulated rooms

The STI for four different impulse responses from different rooms was calculated with Matlab and the results can be seen in table 2.2. The code can be found in the appendix A.

<table>
<thead>
<tr>
<th>Simulated impulse response</th>
<th>STI</th>
</tr>
</thead>
<tbody>
<tr>
<td>auditorium11J02.wav</td>
<td>0.62</td>
</tr>
<tr>
<td>BostonSHJ01.wav</td>
<td>0.44</td>
</tr>
<tr>
<td>GrundvigsJ03.wav</td>
<td>0.33</td>
</tr>
<tr>
<td>Listen04J01.wav</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Table 2.2: STI values for different rooms.
3.1 Spectrum of speech

It is possible to determine the sex of from a long-term magnitude spectrum by looking at the low frequencies of the long-term magnitude spectrum of the speech signals that appear in figures 2.1, 2.2, 2.3 and 2.4. Male have lower fundamental frequencies than females. Fundamental frequencies for males are around 125 Hz whereas females have fundamental frequencies around 250 Hz. [Poulsen, 2005, p. 47]. Figures 2.1 and 2.3 imply female speakers as a peak appears close to 200 Hz, that corresponds to their fundamental frequency. On the other hand, figures 2.2 and 2.4 imply male speakers as the fundamental frequency appears to be around 100 Hz.

Is it very difficult though to determine the language used from a given long-term magnitude spectrum. This is because languages have similar long-term magnitude spectra even though the languages are fundamentally different. This can be seen on the figures appearing in appendix B which show the long-term magnitude spectra for speakers of both genders and different languages.

There is a lot of research going on in the topic of speaker identification. Several techniques and strategies are used and most of them are based on the scheme shown in figure 3.1.

Most of the systems use the dynamic properties of the speech signal rather than analyzing the long-term spectrum. The latter requires a long signal acquisition time to be text-independent. According to figure 3.1 the different blocks can be implemented with different strategies. The features extracted from the speech signal can be the cepstrum and different cepstral coefficients or the coefficients from a linear predictor (LP) based...
on different filter models. The length of the different speech segments is also depending on the strategy. The output of the block after the A/D conversion is a set of vectors (describing different features of speech) that should be compared to the ones stored in a database, corresponding to different individuals, using the concept of distance measurement. Different strategies are used to determine this distance, such as pattern matching based on vector Euclidean distances or the use of Hidden Markov Models (HMM). Finally, a decision device has to determine the individual that provides the maximum likelihood for a given input signal [Campbell, 1997].

3.2 Speech shaped noise

A speech shaped noise generated from a speech signal is shown in figure 2.5. The time fluctuations in the original signal (a), i.e. the envelope, disappear when the phase is randomized (b), while the magnitude spectrum remains unaffected by the process, (c) and (d).

In order to incorporate the fluctuations of the speech signal into the speech shaped noise, two methods are suggested, apart from the ICRA-noises generation described in section 1.3. The first method, illustrated in figure 3.2, uses two inputs: random noise $n(t)$ and a speech signal $s(t)$. The spectrum of the noise is equalized with the one of the speech signal. Using the Hilbert transform, the envelope of the speech is extracted and applied to the equalized noise. The last step is the normalization of the energy, extracted from the spectra of the signal and the modulated speech shaped noise.

Another method is designed to obtain speech shaped noise with modulation features using only a speech signal as shown in figure 3.3. Thus, the noise is obtained by making the phase of the speech random, while the magnitude spectrum is preserved. Then the envelope is applied and the energy is normalized in the same way as in the previous method.

The method shown in figure 3.2 has been implemented in MATLAB (see Analysis_syn-
thesis.m function in appendix [A] and it has been used to produce speech shaped noise with speech envelope from a recorded sound file. The results are shown in figure 3.4. It can be seen that the amplitude of the noise (b) is higher than the amplitude of the speech (a). This is due to the fact that the fine structure of the noise is random noise and the fine structure of the speech is a periodic signal. This last signal carries the same energy as the noise with less amplitude, i.e. the speech has a lower crest factor. It can also be noted that the spectrum of the speech shaped noise is slightly different from the one of the speech signal (c). This is due to the convolution of the original spectrum with the envelope spectrum, which has a reduced bandwidth as can be seen in figure 3.5.

The highest values of envelope magnitude appear at frequencies below 10 Hz, as shown in figure 3.5. It can also be seen that a secondary peak is present around 200 Hz. This peak represents the fundamental frequency of the speaker, in this case a female, which has also been observed in the long-term magnitude spectrum of the same signal in figure 2.1. Therefore, it might be possible to use the envelope spectrum as a feature for speaker recognition.
Figure 3.4: Temporal waveforms of the speech signal (a) and the modulated speech shaped noise (b). The magnitude spectra of both signals and the noise are used as input are shown in (c).

Figure 3.5: Envelope spectrum of the speech signal.
3.3 High-pass and low-pass filtered speech

In order to derive which frequency band is the most important for the intelligibility of speech one can convert the values of AI for the low-pass and high-pass filters, shown in figures 2.8 and 2.9, to the corresponding speech intelligibility values according to figure 1.5 on page 5. The gradient of the resulting curves will be highest at the most important frequency bands for speech intelligibility. This is because, at these frequency regions, small changes in the passing bands of the filters give the highest improvements in speech intelligibility.

In figures 3.6 the dependence of AI and SII on the speech spectrum level can be seen. It can be seen that the AI increases with an increase of the speech level (figures 3.6 (a) and (b)), because the SNR increases. The slopes of the curves are rather similar amongst them and at very low frequencies (125 Hz - 400 Hz) and very high frequencies (5 kHz - 8 kHz) the values of AI, when the speech levels are “loud” and “shout” are identical. On the other hand, the SII presents a quite different behavior. For the low cut-off frequencies of the high-pass filtering it can be seen that the higher the speech level the lower the SII, for cut-off frequencies up to 1.5 kHz. From that cut-off frequency and above the SII is improving as the speech level is increasing. This behavior, which is different to the one observed for the AI concept, is due to the more sophisticated way that the SII procedure is implemented. SII takes into account the different distribution of the energy in the speech spectrum at different levels and the effect of masking.

Thus, taking into consideration figure 3.6 it can be concluded that speech level affects in a different way the two indices.

3.4 Comparison of AI, SII and STI

The idea behind AI (and SII) is that background noise can influence intelligibility due to masking and not all frequency components are equally important. STI is based on the assumption that the envelope of a speech signal must be perceived correctly in order the speech to be understood correctly. This means that AI and SII are calculated directly from the magnitude response of the signals. On the other hand, the results given by STI are correlated with the preservation of the envelope. AI and SII are suited to estimate speech intelligibility of transmission channels in communication systems with limited bandwidth, whereas STI is calculated from the impulse response and is more suited for speech intelligibility in rooms.
Figure 3.6: Effect of different speech levels in the AI (a),(b) and SII (c),(d) for high-pass and low-pass filtering.
STI describes the degradation of intelligibility due to masking in the modulation domain. This correlates with the concept discussed in previous reports about the existence of a modulation filterbank that seems to play an important role in human perception. The idea of a modulation filterbank supports the assumption that STI is based on.
Conclusion

In this report, the intelligibility of speech has been discussed and the main conclusions are the following:

- From the long-term magnitude spectrum, it is possible to determine the gender of the speaker, but not the spoken language. The main peak in the spectrum corresponds to the fundamental frequency, which is around 125 Hz for male speakers and 250 Hz for females.

- Speech-shaped noise is used as a masker for speech intelligibility tests. However, the speech envelope is not always preserved, and some methods have been proposed to generate speech shaped noise with the same speech envelope.

- AI and SII are affected in a different way by the change in the voice level. This is due to the corrections on speech spectrum for different voice levels that SII applies.

- The STI is a useful parameter to assess speech intelligibility in different rooms and its calculation is based in the preservation of modulation in the speech signal. This seems to be in agreement with the way the auditory system perceives sounds and reinforces the assumptions of the existence of a modulation filterbank.
Bibliography


Appendix A

Matlab code

```matlab
1 clc
2 clear all
3 close all
4
5 % Read wav files
6 [y1,fs,Nbits]=wavread('Speech sample 01.wav'); % Female, English
7 [y2,fs,Nbits]=wavread('Speech sample 02.wav'); % Male, English
8 [y3,fs,Nbits]=wavread('Speech sample 03.wav'); % Female, Danish
9 [y4,fs,Nbits]=wavread('Speech sample 04.wav'); % Male, Danish
10
11 % Do something
12 N = length(y1);
13 fres=N ;%fs/N; % Resolution
14 f=fs/fres.*(0:fres/2-1); % Frequency axis
15
16 y1longterm=20*log10(abs(fft(y1,fres)));
17 y2longterm=20*log10(abs(fft(y2,fres)));
18 y3longterm=20*log10(abs(fft(y3,fres)));
19 y4longterm=20*log10(abs(fft(y4,fres)));
20
21 vLout1 = Oct3Smooth(f,y1longterm(1:fres/2),f_AI);
22 vLout2 = Oct3Smooth(f,y2longterm(1:fres/2),f_AI);
23 vLout3 = Oct3Smooth(f,y3longterm(1:fres/2),f_AI);
24 vLout4 = Oct3Smooth(f,y4longterm(1:fres/2),f_AI);
25
26 figure;
27 semilogx(f,y1longterm(1:fres/2),'Color',[.6 .6 .6])
28 hold on;
29 semilogx(f_AI,vLout1,'k','LineWidth',2);
30 xlim([100 10000])
31 ylim([-40 70]);
32 set(gca,'XTick',[100 250 500 1000 2000 4000 8000]);
33 nicefigure;
34 legend('Long term spectrum','1/3-octave smoothed spectrum','Location','SouthWest');
35 xlabel('Frequency [Hz]');
36 ylabel('Magnitude [dB]');
```
% printlatex('EnglishFemale',11,10,'nofigcopy');
figure;
semilogx(f,y2longterm(1:fres/2),'Color',[.6 .6 .6])
hold on;
semilogx(f_AI,vLout2,'k','LineWidth',2);
xlim([100 10000])
ylim([-40 70]);
set(gca,'XTick',[100 250 500 1000 2000 4000 8000]);
nicefigure;
legend('Long term spectrum','1/3-octave smoothed spectrum','Location','SouthWest');
xlabel('Frequency [Hz]');
ylabel('Magnitude [dB]');
% printlatex('EnglishMale',11,10,'nofigcopy');
figure;
semilogx(f,y3longterm(1:fres/2),'Color',[.6 .6 .6])
hold on;
semilogx(f_AI,vLout3,'k','LineWidth',2);
xlim([100 10000])
ylim([-40 70]);
set(gca,'XTick',[100 250 500 1000 2000 4000 8000]);
nicefigure;
linspace('Long term spectrum','1/3-octave smoothed spectrum','Location','SouthWest');
xlabel('Frequency [Hz]');
ylabel('Magnitude [dB]');
% printlatex('DanishFemale',11,10,'nofigcopy');
figure;
semilogx(f,y4longterm(1:fres/2),'Color',[.6 .6 .6])
hold on;
semilogx(f_AI,vLout4,'--','Color',[.6 .6 .6])
semilogx(f_AI,vLout2,'k');
semilogx(f_AI,vLout3,'--k');
xlim([100 10000])
ylim([20 60]);
set(gca,'XTick',[100 250 500 1000 2000 4000 8000]);
nicefigure;
legend('English female','English male','Danish female','Danish male','Location','SouthWest');
xlabel('Frequency [Hz]');
ylabel('Magnitude [dB]');
% printlatex('SpecComparison','11,10','nofgsocpy');

% r_y1 = randomizePhase(y1);
% r_y2 = randomizePhase(y2);
% r_y3 = randomizePhase(y3);
% r_y4 = randomizePhase(y4);

clear all;
clear variables;

[y,fs,Nbits]=wavread('Speech sample 01.wav'); % Female, English
N = length(y);
t=1/fs*(1:N);
f=fs/N.*(0:N/2-1); % Frequency axis
yfft=fft(y);
ylongterm=20*log10(abs(yfft));
vLout = Oct3Smooth(f,ylongterm(1:N/2),f_AI);
y_energy = sum(10.^(-vLout/10));

y_envelope = abs(hilbert(y));

figure;
plot(t,y,'Color',[.5 .5 .5]);
hold on
plot(t,y_envelope,'k','LineWidth',2);

x=rand(N,1);
figure; plot(t,x,'Color',[.5 .5 .5]);
x=x.*y_envelope;
hold on
plot(t,x,'k');

xfft=fft(x);
xlongterm=20*log10(abs(xfft));
xvLout = Oct3Smooth(f,xlongterm(1:N/2),f_AI);
figure(111);
semilogx(f_AI,xvLout,'--k');
hold on
xfft=xfft.*abs(yfft)./abs(xfft);
x2longterm=20*log10(abs(xfft));
x2vLout = Oct3Smooth(f,x2longterm(1:N/2),f_AI);
semilogx(f_AI,x2vLout,'-k');
x2=ifft(xfft);
x2=x2.*y_envelope;
x2fft=fft(x2);
x2longterm=20*log10(abs(x2fft));
x2vLout = Oct3Smooth(f,x2longterm(1:N/2),f_AI);
x2_energy = sum(10.^(-x2vLout/10));
x2 = x2.*sqrt(y_energy/x2_energy);
figure;
plot(t,x2,'k');
xlabel('Time [s]');
xlim([0 50]);
ylim([-4 4]);
nicefigure;
printlatex('noise2_time',7,8,'nofigcopy');

figure(111);
semilogx(f_AI,x2vLout+10*log10(y_energy/x2_energy),'--','Color',[.5 .5 .5]);
xlim([100 100000]);
xlabel('Frequency [Hz]');
ylim([20 55]);
ylabel('Magnitude [dB]');
nicefigure;
legend('Input noise','Speech','Output noise','Location','SouthWest');
printlatex('specs_mod_shape',9,8,'nofigcopy');
Appendix B

Long-term average spectrum of speech

![Graph showing the long-term average spectrum of speech for different regions and categories.]

Figure B.1: Long-term average spectrum of speech. From [Byrne et al., 1994, p. 2113].
APPENDIX

Figure B.2: Long-term average spectrum of speech. From [Byrne et al., 1994, p. 2114].
FIG. 3. Male and female LTASS values for Cantonese, Mandarin, Vietnamese, Japanese, and French. Solid line shows LTASS average across 17 speech samples.

Figure B.3: Long-term average spectrum of speech. From [Byrne et al., 1994, p. 2115].

FIG. 4. Male and female LTASS values for Singhalese and Welsh. Solid line shows LTASS average across 17 speech samples.

Figure B.4: Long-term average spectrum of speech. From [Byrne et al., 1994, p. 2116].