Statistical analysis of speech signals

Blomberg, M. and Elenius, K. O. E.

journal: STL-QPSR
volume: 11
number: 4
year: 1970
pages: 001-008

http://www.speech.kth.se/qpsr
I. SPEECH ANALYSIS

A. STATISTICAL ANALYSIS OF SPEECH SIGNALS

M. Blomberg and K. Elenius

This is a condensed report of a thesis study carried out at the Department of Speech Communication in 1970. The purpose was to determine, for continuous speech, peakfactor, formfactor, long-time average spectrum of voiced and voiceless sections separately, spectral density at different voice intensity levels, distribution of the speech-wave amplitude, statistics on pause lengths and long-time average RMS of the speech wave. All tasks have been solved using the CDC computer of the Department.

Speech Material

Sixteen 1-min long sections of conversational speech were recorded with the participants on separate channels of a twin-track tape-recorder. The recording was made by the Swedish Board of Telecommunications (Televerket), Farsta. A limitation of this material was the relative low degree of cross-channel attenuation, 25-30 dB. This distortion was found to impair the statistics on pause lengths only. For this particular study we recorded another material with higher inter channel attenuation.

Distribution of Speech-wave amplitude

The amplitude distribution of speech was measured with the equipment shown in Fig. 1-A-1. The computer samples in real time the rectified and short-time (13 msec) integrated speech wave and assembles the distribution by adding 1 to the word in storage that corresponds to the quantized level of the sample. The distribution is displayed as a histogram with a resolution of 2 dB. Fig. 1-A-2 shows a typical histogram of one of the samples of conversational speech, earlier described. The amplitude of the curve is a measure of the probability that the signal level will have the value given by the x-axis.

The curve shows two distinct maxima. The one for higher signal levels depends on the speaker that is to be measured. The one for lower levels is due to noise and the voice of the speaker on the other channel. The distance between the maxima gives a measure of the signal-to-noise ratio, or as in this case, the attenuation between the channels which evidently is 25-30 dB. An exact measure is not possible, as the long-time RMS of the speakers cannot be derived from the histogram.
Fig. I-A-1. The equipment used to measure the distribution of the speech-wave amplitude.
Fig. I-A-2. Histogram of the distribution of the speech-wave amplitude. Conversational speech.
For appropriate measurements on pause lengths, it is of importance where the level of the pause threshold is situated. It should be set as low as possible without allowing noise to exceed the threshold. Fig. I-A-3 shows an amplitude histogram where only one speaker is talking. The noise density maximum is about 30 dB under the speech density maximum. The choice of the pause threshold level should be ca 15-20 dB under the speech density maximum, where the histogram has a minimum. The total pause length is then least dependent of the threshold level. This value is about 35 dB less than the highest speech level during the measured interval.

Pause-length statistics

There are two essential parameters that define the acoustic boundaries of a pause, namely, maximum amplitude within the pause and its minimum duration. The former puts an upper limit to the range of speech-wave amplitudes which might comprise a pause. The latter fixes the smallest amount of time during which these acoustic amplitudes must flow in order to continue a pause. Otherwise inter-wave "gaps", that is, moments of relatively less amplitude that usually are present in a complex speech wave, might be detected as pauses. The duration of the inter-wave "gap" is a very short one and depends on the fundamental frequency, the particular shape of the wave and the level of the pause-maximum amplitude. It is reasonable to state that for complex waves of fundamental frequency of 100 Hz the duration of the inter-wave "gaps" is necessarily shorter than the period, 10 msec. Pauses, or most pauses, last more than these durations$^4$. The sampling rate of the rectified speech wave was 1600 Hz. This should be enough to exclude inter-wave "gaps".

Automatic pause-length analysis is of course effected by clicks, gurgitations, air-stream noise, and other assorted sounds. These sounds will sometimes exceed the amplitude threshold and will, at worst, ripple just about the threshold level causing several short pauses (usually 10 to 30 msec) in a speech segment, which otherwise would be considered as being one pause, since it adds nothing to the speech intelligibility. Raising the pause-maximum amplitude to eliminate these effects will cause the weakest consonants to be considered as pauses. However, for pause lengths longer than approximately 30 msec the pause statistics most probably differ little from those measured with the effects above eliminated.
Fig. I-A-3. Histogram of the distribution of the speech-wave amplitude. Only one speaker.
The distribution of pause lengths has been measured for pauses between 12.5 and 250 msec and for pauses up to 1 sec separately. The longer pauses were measured after integrating the speech wave ($\tau = 13$ msec). This made it possible to derive the pause-maximum amplitude from the distribution function of the integrated speech wave as mentioned above. The threshold level was about 35 dB lower than the maximum intensity.

Pauses shorter than 250 msec were measured directly from the rectified speech wave (no integration) using a pause-minimum duration time of 12.5 msec. Fig. I-A-4 shows the distribution and the cumulative distribution function of 375 pauses, shorter than 250 msec, measured in three conversations, each about 1 min long. The pause maximum amplitude was 30-34 dB lower than the rectified (not integrated) average peak amplitude in each conversation. The distribution was little affected by this parameter. Raising it by 16 dB made the cumulative distribution function vary less than 10%, but it increased the number of pauses by 15%. The cumulative distribution function can be approximated by an exponential:

$$F = 1 - \exp[-(t-12.5)/\lambda]$$

$\lambda = 41$ msec

12.5 msec $< t < 250$ msec

The deviation is less than 5% for the cumulative distribution function (for pauses longer than 75 msec it is less than 2%). The average of this approximation is 53.5 msec ($\lambda + 12.5$). This is very close to the measured average, which is 54 msec.

The distribution of pause lengths up to 1 sec is shown in Fig. I-A-5. Pause lengths longer than 300 msec can be described by a rectangular distribution. This will give the following cumulative distribution function:

$$F = 0.92 + 0.11(t-300)/1000$$

300 $< t < 1000$ msec

The deviation is less than 1%. These pauses contain breaks due to thinking and breathing and due to prosody. Of course, these pauses may be shorter (150 msec), while occlusions seldom are longer than 100 msec for continuous speech.
Fig. I-A-4.  Pause-length statistics.  12.5 - 250 msec.
Fig. I-A-5. Pause-length statistics. 12.5 - 1000 msec.

\[ F = 0.92 + 0.11 \frac{t - 300}{1000} \]
The long-time mean value of the rectified speech wave (LMV) was measured by averaging the rectified and integrated speech signal. Together with the formfactor ($f$) it will give the long-time average RMS ($LRMS$), since

$$f = \frac{LRMS}{LMV}$$

In order to measure the LMV for speech only and not for all pauses, the following criterion has been used. If the level of the rectified and integrated speech wave is lower than a given threshold for a shorter time than 100 msec it will be used in the calculation of LMV, while pauses longer than 100 msec will be neglected. The boundary was chosen at 100 msec after studying occlusion lengths in continuous speech. Later measurements showed that this parameter was not critical. The same results were obtained using the boundaries 50 msec and 500 msec. Likewise the other parameter, the threshold level, was not critical either. Changing it by 20 dB did not effect the result which, however, was quantized in steps of 2 dB. The value of $LRMS$ was later used as a reference level when a speech controlled, loud-speaking telephone system was simulated by the computer. The results of this investigation will be discussed in a forthcoming report.

Formfactor

The formfactor was derived in the following way. The speech signal was sampled and digitalized. In real time, the computer accumulates samples $x_n$ and calculates the sums $\sum x_n$ and $\sum x_n^2$. The formfactor is then computed from the equation:

$$f = \frac{\sqrt{\sum x_n^2}}{1/\sum |x_n|}$$

With this definition the formfactor of a sinc wave will be 1 dB. Pauses in speech were treated the same way as when calculating LMV (see above).

The sampling frequency used was 10 kHz. This is somewhat too low for securing a complete waveform reconstruction. However, it can be shown that the formfactor can be derived from the statistical distribution of the signal level. The problem is then reduced to choosing a sampling density that allows a sufficiently good approximation to the distribution. This condition is easier to meet.
Measurements on conversational speech of 1 min length varying the sampling frequency between 50 and 10,000 Hz have verified this. See Table I-A-1.

Table I-A-1. Formfactor at different sampling frequencies. The formfactor of each speaker was determined. The results are shown in Table I-A-2.

<table>
<thead>
<tr>
<th>Sampling frequency</th>
<th>50</th>
<th>500</th>
<th>5000</th>
<th>10000 Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formfactor</td>
<td>1.4</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Table I-A-2. Distribution of formfactors

<table>
<thead>
<tr>
<th>Formfactor</th>
<th>3 dB</th>
<th>4 dB</th>
<th>5 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of male subjects</td>
<td>1</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>&quot; &quot; female &quot;</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Total number</td>
<td>2</td>
<td>13</td>
<td>1</td>
</tr>
</tbody>
</table>

The formfactor varies between 3 and 5 dB with a mean value of about 4 dB. This corresponds well to measurements by Fant(3), who found a value of 3.5 dB.

Peakfactor

The peakfactor, defined as

$$\rho = \frac{\text{largest value during the observed time}}{\text{RMS}}$$

has been measured for 1 min long parts of speech and 10 sec long sections of the same speech. The results give a peakfactor of 16.5 ± 1.0 dB for the 1 min long parts, while the corresponding value for the sections of 10 sec was 17.0 ± 1.5 dB. The small difference indicates that the peakfactor does not vary much with durations longer than 10 seconds.

Fant(1) and Sjöholm(5) have measured a peakfactor of 17 dB, which conforms quite well with these results.
**Long-time average speech spectrum**

The 51-channel filterbank was used for measurements of long-time average speech spectrum, including separate statistics for voiced and voiceless segments. The computer program involved a calculation of the mean value of sampled spectrum sections.

The total range of the filterbank was 0-9.6 kHz. Constant filter bandwidths of $B = 250$ Hz and a mel-scale frequency display was chosen. The detection of voiced and voiceless segments of speech employed one channel with normal high-frequency pre-emphasis (6 dB/oct, 200-5000 Hz) and one with low-pass filtering (330 Hz) of the signal, see Fig. I-A-6. Sound is detected when the output level of the pre-emphasized signal is higher than a threshold level, located ca 30 dB below the maximum intensity during the measured interval. This is not exactly comparable, though, to the pause-maximum amplitude level mentioned above, as the weaker voiceless fricatives will be stronger because of the pre-emphasis. If the output of the LP-filtered signal is higher resp. lower than an adjusted level, the sound is judged as voiced resp. voiceless. When the condition for the sound class that is to be measured are met, the filters of the filterbank are synchronously sampled, and the spectrum sections are added to an accumulative area in the computer storage to form the average spectrum. Maximum sampling rate is about 160 Hz.

The effects of the low cross-channel attenuation were eliminated by means of sound detection on the other channel (channel 2) than the one measured. When sound is detected no measurement is made. Otherwise, voiced segments of the speaker on this channel might be detected as voiceless on channel 1 and give a wrong average spectrum of voiceless segments.

Average spectra with no separation voiced-voiceless for one male and one female speaker that have been computed by this method are shown in Fig. I-A-7 and I-A-8. They are compared with spectra after Dunn & White\(^{(7)}\) and French & Steinberg\(^{(8)}\). There is no great difference to be found.

Separate spectra of voiced and voiceless segments of the same speakers are shown in Figs. I-A-9 and I-A-10. Voiced sounds, being more frequent and stronger than the unvoiced, dominate the spectrum of the total speech wave. Especially, this is the case for frequencies below 3 kHz. Over 3 kHz spectrum falls more rapidly.
Fig. I-A-7. Long-time average spectrum. Male subject, M1.
Fig. I-A-8. Long-time average spectrum. Female subject, K4.
The voiceless sounds, which are characterized by lack of glottal pulses, include not only voiceless fricatives but also aspiratives, bursts after stops, and, more or less desirable, breathing sounds before sentences. These sounds and dark /sj/ will contribute to a spectrum shape deviating from the expected high-frequency emphasis.

Fig. I-A-11 shows an average spectrum for the aspirative, whispered vowels æ₁, ø₁, u₁, ð₁, œ₁, i₁, y₁, æ₂, and ø₂. The energy is mainly concentrated below 2 kHz. This finding is in agreement with the average spectrum of voiceless sounds in this region, because the aspiratives constitute a great part of these.

Spectra of voiceless sounds have an energy minimum between 3 and 4 kHz. Most probably, this depends on the fact that neither the aspiratives nor the fricatives contain any larger amount of energy at this frequency. The fricatives, where /s/ is the most frequent, have the main part of their energy above 4 kHz and the aspiratives below 2 kHz.

**Average spectrum at different voice intensities**

Two subjects made a 20-sec long utterance, the same for both, 4 times, varying the voice intensity in 10 dB steps between -10 and +20 dB rel. normal speech as measured by a VU-meter. Average spectra without voiced-voiceless separation were measured. The recording and playback levels were adjusted so that low-voice effort would give about the same output signal as high. This was done in order to get the same precision for all intensities.

The spectra are shown in Fig. I-A-12 and I-A-13. High-voice effort differs from low mainly by the spectrum slope below 1 kHz. For speaker J.S., the intensity falls from 300 to 1000 Hz ca 10 dB for -10 dB voice intensity and only ca 5 dB for +20 dB. The corresponding values for speaker K.E. are 35 resp. 10 dB. This effect is due to the voice source. At low-voice effort, source spectrum falls more rapidly in this frequency region\(^2\). The level in the fundamental frequency region is higher for low-voice intensity, which is an apparent feature of the figures. A combining effect is also the increase of F0 at higher voice intensities.

For medium and high-voice effort, there is a distinct peak between 3 and 4 kHz. This is the so-called "singing formant". It is related to the sinus piriformis, a cavity around the upper part of larynx. It is enlarged when the
Fig. I-A-11. Average spectrum. Whispered vowels.
Fig. I-A-12. Average spectrum at different voice effort. Male speaker J.S.
Fig. I-A-13. Average spectrum at different voice effort. Male speaker K.E.
voice intensity increases, which probably causes new formants to be formed compared with normal speech. A more thorough explanation is made by J. Sundberg in this issue of STL-QPSR.

References:
(3) G. Fant: "Undersökning av 10 sek standardfras", LM Ericsson protokoll nr H/P-1051.