Moment weighting techniques for segmentation

Liljencrants, J.

journal: STL-QPSR
volume: 3
number: 4
year: 1962
pages: 015-021

http://www.speech.kth.se/qpsr
C. MOMENT WEIGHTING TECHNIQUES FOR SEGMENTATION

The automatic segmentation of different speech sounds depends on a suitable choice of parameters, some of which can be extracted from the signal by means of moment weighting. The moments may be evaluated from a spectral analysis carried out for instance with a filter bank. From the technical point of view it is, however, less involved to use direct methods for moment measuring.

In this study we are concerned with the concepts of center of gravity and spread in the spectra of some unvoiced fricative sounds. The methods are of course applicable to other sound classes as well.

Considering a spectrum with the power density \( S(f) \) we have the moment of order \( n \) around the origin as

\[
M_n = \int_{0}^{\infty} f^n \cdot S(f) \cdot df
\]

The center of gravity of the spectrum is

\[
T = \frac{M_1}{M_0}
\]

The spectral spread is the mean deviation from \( T \) given by

\[
s = \sqrt{\frac{M_2}{M_0} - T^2}
\]

where \( M_2 \) is the central moment of order two found from

\[
M_2 = \int_{0}^{\infty} (f - T)^2 \cdot S(f) \cdot df
\]

The total level of the signal is \( M_0 \). If the signal is passed through an emphasis network with the slope +3 dB/oct the power density will be multiplied with a factor that is proportional to frequency. The total level of the processed signal will thus be proportional to \( M_1 \). When the signal is processed in an integrating network with negative slope the resulting total level will be proportional to a moment of negative order.

Moments around a frequency \( f_x \) might be approximated if the signal is first modulated with \( f_x \) in a balanced modulator. The resulting low frequency components could then be treated as above. Moments around \( f_x \) with negative orders are of course out of question.
if the signal contains energy at $f_x$ in the frequency domain.

In the following experiment various moments have been approximated by measuring the mean average value of the processed signal instead of the RMS value. The errors introduced by this should be negligible.

It soon was found that the low order moments were too insensitive to spectral changes for practical purposes when a reasonable accuracy in the measuring equipment is assumed. This sensitivity may be increased by the use of higher order moments. Then we cannot utilize the conventional definitions of center of gravity and spread. These quantities may instead be interpreted by means of a simple model:

Assume a rectangular (noise) spectrum covering the band from $f_1 = f_c/\alpha$ to $f_2 = f_c \cdot \alpha$ and with constant power density $S_o$. $f_c$ will then be the center of gravity if a logarithmic frequency scale is assumed. The relative spread is represented by the factor $\alpha$. The moments of this signal are given by

$$
\begin{align*}
M_{n-1} &= S_o \cdot \frac{f_{c}^{2n} - 1}{\alpha^{2n}} ; & n &\neq 0 \\
M_{-1} &= S_o \cdot 2 \cdot 1 \cdot \alpha
\end{align*}
$$

(5)

It is now desirable to find two simple combinations of moments giving measures of $f_c$ and $\alpha$ respectively. Furthermore these combinations should be level normalized (independent of $S_o$).

$f_c$ is most easily found from expressions of the type

$$
\frac{M_{n+v}}{M_n} = f_c^{2n} \cdot F(\alpha) 
$$

(6)

where $F(\alpha)$ is a function of $\alpha$. The special case $n = 0, v = 1$ is equal to formula (2). Some of these expressions are independent of $\alpha$, indeed

$$
\frac{M_{n-1}}{M_{n-1}} = f_c^{2n}
$$

(7)

For practical purposes it seems to be justified to consider moments of orders not higher than $\pm 4$. We are then left with three non-
trivial cases of formula (7), namely
\[
\begin{align*}
\frac{M_2}{M_{-2}} &= \lambda_2^2, \\
\frac{M_4}{M_{-4}} &= \lambda_4^2, \\
\frac{M_2}{M_{-4}} &= \lambda_6^2
\end{align*}
\]
which vary 6, 12, and 18 dB respectively per octave shift in \(f_0\).
Looking for combinations sensitive to \(\lambda\) only a brief investigation indicates that symmetrical expressions of the type \(M_n \cdot M_{-n}/M_o^2\) tend to be optimal. Two combinations of this type are plotted on logarithmic scales in fig. I - 7. (When the logarithmic measures dB and octaves are used the variable is put between square brackets.)

A technically different method to establish moment-related quantities in noise spectra is the measurement of the short time average zero crossing rate of the time domain signal. This is given by Rice (1) as
\[
\rho_o = 2 \sqrt{\frac{M_2}{M_o}}
\]
(9)
If the signal is emphasized \(3 \cdot n\) dB/oct prior to the zero crossing measurement it follows that
\[
\rho_n = 2 \sqrt{\frac{M_{n+2}}{M_n}}
\]
(10)
Treating the signal in a differentiating network (+6 dB/oct) or integrating (-5 dB/oct) will make it possible to derive \(\rho_2\) and \(\rho_{-2}\) from which we can get for instance
\[
\begin{align*}
\rho_2^2 &= 4 \cdot \frac{M_o}{M_{-2}} = 4 \cdot \lambda_2^2 \\
\frac{\rho_{-2}}{\rho_2} &= \sqrt{\frac{M_4}{M_o} \cdot \frac{M_{-4}}{M_2}} \approx 0.77 \lambda \quad \text{if, say} \lambda \approx 1.5
\end{align*}
\]
(11)
The measures discussed above are inherently normalized. It is possible to use simpler combinations of moments to derive corresponding measures if the input signal is passed through an AGC amplifier. The requirements on an amplifier of this kind are however very strict. The time constants in its regulating circuits must be sufficiently short to equalize the level variations
Fig. 1. The sensitivity of some moment combinations to $\lambda$. 

\[
\begin{align*}
\frac{\rho_2}{\rho_{-2}} \\
\frac{M_4 M_{-4}}{M_0^2} \\
\frac{M_2 M_{-2}}{M_0^2}
\end{align*}
\]
within individual sound segments. In addition the output level variations must be very small. The use of a less sophisticated compressor will however ease the requirements on dynamic range in the following circuitry.

The normalization does not exist with the zero crossing method. Great care has however to be exercised in the design of the clipping stages to insure high sensitivity and freedom from blocking phenomena.

**Measurements**

Measurements have been made on a small material from a phonetically trained speaker*. The fricatives s, f, š, and ž appeared in initial position in the Swedish words of the type /Co:na/. Four utterances of each fricative were used. The gross spectral features of the speech material are displayed in Fig. I - 8.

The signal was band limited to the range from 0.5 to 12 kc/s with moderately sharp filters (180 dB/octave). The signal was then fed into five networks giving -12, -6, -0, +6, and +12 dB/oct slope. The signals from the emphasis networks were rectified, smoothed (integration time constant 25 ms), and synchronously sampled at 23 ms intervals. The subsequent processing stages providing analog-digital conversion on a logarithmic scale have been described in ref. (2).

Results of the measurements are plotted in Fig. I - 9. Points belonging to the same utterance of a specific sound are joined with arrows pointing in the direction of time.

The figure displays a certain correspondence to the distinctive features "compact/diffuse". The spectrum model could be expected to match "compact" sounds better than others. This seems to be verified by the listening test reported on in the appendix.

---

*Part of the material used by Mártony (3).
Fig. 1 - 8. Mean spectra of five Swedish fricative sounds. Each diagram represents four utterances. Each utterance has been sampled at an average five points 10 ms apart. Each sample is normalized with respect to its total level which is used as reference. 15 per cent of the samples fall between each pair of lines.

Analysis was performed by contiguous third-order Butterworth filters. The effective averaging time constant of the smoothing filters is of the same order as the sampling interval. Only samples with a total level within 50 dB (for /f/ 15 dB) from the maximum for each utterance have been considered.
Fig. 1-9. Moment weighting results for four utterances of each of five sounds. Sloping dashed lines indicate loci of $f_c$. $(M_4/M_{-4})$ is not independent of $X$. 
Po, and P-2 was experimentally determined for the same speech material. \( f_c \) and \( \Lambda \) were determined from these and conform in quality to a reasonable extent with the previous results. However, it appeared that the zero crossing method was apt to give a less distinct separation between the different sounds. This seems to be due to hysteresis phenomena in the clipping process.

It should be observed that some signals may give rise to values of \( \kappa^2 \) that are greater than the ratio between the limiting frequencies of the entire processed band. This will occur when the signal spectrum has two widely separated peaks of comparable magnitude. The /f/ sound tends somewhat to this. Such a signal will of course match the rectangular spectrum model extremely bad. The effect may also be disturbing in case of excessive background noise or hum and great care should be taken in case multiple integration or differentiation is used.

Appendix

In order to partially justify the model a set of sounds were fabricated by filtering out certain bands of a random white noise signal according to the specification below. 180 dB/oct filters were used for this purpose. The data approximately correspond to mean values obtained from the measurements above.

<table>
<thead>
<tr>
<th>sound</th>
<th>( f_1 )</th>
<th>( f_2 )</th>
<th>Total level</th>
<th>( f_m )</th>
<th>( \kappa^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f )</td>
<td>0.5</td>
<td>8.0</td>
<td>-25</td>
<td>2.0</td>
<td>4.0</td>
</tr>
<tr>
<td>( s )</td>
<td>4.0</td>
<td>8.0</td>
<td>-10</td>
<td>5.6</td>
<td>1.0</td>
</tr>
<tr>
<td>( \varphi )</td>
<td>2.0</td>
<td>6.0</td>
<td>-15</td>
<td>3.5</td>
<td>1.7</td>
</tr>
<tr>
<td>( \phi )</td>
<td>1.5</td>
<td>4.0</td>
<td>-15</td>
<td>2.5</td>
<td>1.4</td>
</tr>
<tr>
<td>( \phi )</td>
<td>0.75</td>
<td>3.0</td>
<td>-20</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>background</td>
<td>0</td>
<td>16</td>
<td>-35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Samples with the duration 180 ms were spliced in lieu of the /f/ in identical copies of the natural utterance /fe:na/. The maximum level of this word is the reference in the table above. There was apparently no formant transition at the beginning of the /e:/. The onset and decay of the synthetic sounds lasted some 6 ms controlled by diagonal tape splicings. The samples were assembled in random order into two lists of 4x5 items and background noise was added. The stimuli were presented over headphones to nine members of the laboratory staff for forced choice identification. Three subjects were phonetically trained and responded 82 % "correct", somewhat higher than the average 74 %. The confusion matrices below are pooled for all subjects. The first and the second lists are kept apart to show the learning. Separation between /ʃ/ and /ɻ/ was requested although there is no semantical difference between them in Swedish.

<table>
<thead>
<tr>
<th></th>
<th>f</th>
<th>s</th>
<th>ʃ</th>
<th>ɻ</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>22</td>
<td>13</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ʃ</td>
<td>1</td>
<td>23</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>ɻ</td>
<td>2</td>
<td>4</td>
<td>23</td>
<td>5</td>
</tr>
<tr>
<td>ʃ</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>f</th>
<th>s</th>
<th>ʃ</th>
<th>ɻ</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>28</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>s</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ʃ</td>
<td></td>
<td>27</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>ɻ</td>
<td>2</td>
<td>4</td>
<td>28</td>
<td>2</td>
</tr>
<tr>
<td>ʃ</td>
<td></td>
<td>11</td>
<td>2</td>
<td>23</td>
</tr>
</tbody>
</table>

The good separation between /ʃ/ and /ʃ/ is somewhat surprising having the study of Mártiny (3) in mind but could be explained by that these particular stimuli had a greater difference in $f_0$ than found in the live speech. The /ɻ/ sound was very unnatural partly due to the lack of formant transitions in the following vowel. It resembled a "telephone quality" /ʃ/. Some subjects offered quality judgements which essentially consisted with the ranking in the matrix diagonals.

J. Liljencrants
References

