Filterbanks Used in FMS

This document provides full mathematical specifications for the two filter banks used in FMS. We consider this document important for reproducibility. (It is true that the Python and Matlab code also reproduce our results, but mathematics is the *universal* langauge.) Since this level of detail is not suitable for a four-page paper, we have elected to post it with our software. For anonymity, we are not advertising this location at this time, so we are instead including this document in the submission as supplemental material.

I. Mel Spectrum Filterbank

The mel-spectrum filterbank used to create the fixed-size modulation spectrum (FMS) accumulates Hertz-scale spectral samples to create a mel-scale spectral representation [1]. The filters have uniform width and spacing on the mel scale and triangular shape on the Hertz scale. There are N_{mel} filters $\theta_{k,i}$, $i = 0$ to $N_{mel} - 1$, $k = 0$ to $N_{Hertz} - 1$. Since the length of the DFT that produces the Hertz-scale spectral representation is N_t (even), it follows that $N_{Hertz} = (N_t/2) + 1$.

The upper frequency limit for the analysis is f^u (in Hz). This value is converted to a mel-scale value $\tilde{f}^u$ (in mel) using the relationship given in [1],

$$\tilde{f}^u = 2595 \log_{10}(1 + f^u/700), \tag{1}$$

and the resulting range is evenly divided:

$$\tilde{\Delta} = \tilde{f}^u/(N_{mel} + 1). \tag{2}$$

Let $\tilde{b}_i = i\tilde{\Delta}$, for $i = 0$ to $N_{mel} + 1$. Then the filter centered at $\tilde{b}_i$ ($i = 1$ to N_{mel}) extends from $\tilde{b}_{i-1}$ to $\tilde{b}_{i+1}$. We use the inverse of (1) to convert the $\tilde{b}_i$ to their Hertz scale equivalents b_i:

$$b_i = 700 \left(10^{(\tilde{b}_i/2595)} - 1\right). \tag{3}$$

The filterbank values are given by

$$\theta_{k,i} = \begin{cases} \eta_i \frac{f_k - b_i}{b_{i+1} - b_i}, & b_i \le f_k < b_{i+1}, \\ \eta_i \left(1 - \frac{f_k - b_{i+1}}{b_{i+2} - b_{i+1}}\right), & b_{i+1} \le f_k < b_{i+2}, \\ 0, & \text{otherwise}, \end{cases}$$
$$k = 0 \text{ to } N_{Hertz} - 1, \qquad i = 0 \text{ to } N_{mel} - 1. \tag{4}$$

The Hertz-scale frequencies f_k used in (4) are calculated from the audio sample rate f_s and the DFT length N_t (even):

$$f_k = k\frac{f_s}{N_t}, \quad k = 0 \text{ to } N_t/2. \tag{5}$$

The limits for k in given in (4) and (5) are consistent because $N_{Hertz} = (N_t/2) + 1$. Normalization for the width of each band is accomplished by

$$\eta_i = \frac{1}{b_{i+2} - b_i}, \quad i = 0 \text{ to } N_{mel} - 1. \tag{6}$$

Using (4) we can show that the upper slope of filter $\theta_{\cdot,i-1}$ and the lower slope of filter $\theta_{\cdot,i}$ intersect at

$$\frac{\eta_{i-1} b_{i+1} + \eta_i b_i}{\eta_{i-1} + \eta_i} \text{ Hz}. \tag{7}$$

In the simplified case $\eta_{i-1} = \eta_i$, the intersection given in (7) becomes the midpoint of b_i and b_{i+1}. We can also use (4) to show that the filter centered at b_i has a peak value of η_i and at the simplified intersection locations $(b_i + b_{i-1})/2$ and $(b_i + b_{i+1})/2$ it has value $\eta_i/2$. Thus, at the simplified intersection location, the value of the filter relative to its peak value is $1/2$ or -6.0 dB.

II. Modulation Spectrum Filterbank

The modulation spectrum filterbank accumulates N linear-Hertz-scale spectral samples to create a logarithmic-Hertz-scale spectral representation with N_{mod} samples. When viewed on a logarithmic frequency scale, the filters have uniform spacing and identical triangular shape. Following the successful precedent established in [2], the first filter is centered at 4 Hz and the last filter is centered at 128 Hz, so it is required that $2 \le N_{mod}$. The filterbank contains N_{mod} filters $\Phi_{k,m}$, $m = 0$ to $N_{mod} - 1$, $k = 0$ to $N - 1$. Since the number of per-frame results used in the modulation spectrum DFT is N_f, the number of resulting unique spectral values is $\lfloor N_f/2 \rfloor + 1$. In other words, the number of linear-Hertz-scale spectral values used as input to the filterbank is $N = \lfloor N_f/2 \rfloor + 1$.

The modulation spectrum filterbank is constructed as follows. The range from 4 to 128 Hz is evenly divided on the log scale:

$$\bar{\Delta} = \frac{\log_2(128) - \log_2(4)}{N_{mod} - 1}. \tag{8}$$

The initial log-scale center frequencies are given by

$$\bar{b}_m = \log_2(4) + m\bar{\Delta}, \quad m = 0 \text{ to } N_{mod} - 1. \tag{9}$$

To maximize consistency across different file lengths, we move each of these initial filter center frequencies to match the nearest DFT bin. The spacing of the DFT bins can be calculated from the audio sample rate f_s, the frame stride N_s (number of samples of advance when forming the next frame) and the number of frames used in the modulation spectrum DFT N_f:

$$\Delta_h = \frac{f_s}{N_s N_f}, \tag{10}$$

and the log frequencies of the DFT bins are given by

$$\bar{f}_k = \log_2(k\Delta_h), \quad k = 0 \text{ to } N - 1, \tag{11}$$

with $\log_2(0)$ defined to be $-\infty$. Moving $\bar{b}_m$ to match the nearest value of $\bar{f}_k$ is achieved by

$$\bar{b}'_m = \log_2\left(\left[\frac{2^{\bar{b}_m}}{\Delta_h}\right]\Delta_h\right), \quad m = 0 \text{ to } N_{mod} - 1, \qquad (12)$$

where $[\cdot]$ indicates rounding to the nearest integer. Due to the logarithmic scale, these adjustments are very small near the top of the frequency scale (128 Hz) but larger near the bottom of the scale (4 Hz). The filter half-widths are given by

$$\bar{\delta} = \bar{\Delta}/(2 - \sqrt{2}), \qquad (13)$$

and these values were selected to make unweighted adjacent filters intersect at -3 dB.

The filterbank values are given by

$$\Phi_{k,m} = \begin{cases} \nu_m \frac{\bar{f}_k - (\bar{b}'_m - \bar{\delta})}{\bar{\delta}}, & \bar{b}'_m - \bar{\delta} \le \bar{f}_k < \bar{b}'_m, \\ \nu_m \left(1 - \frac{\bar{f}_k - \bar{b}'_m}{\bar{\delta}}\right), & \bar{b}'_m \le \bar{f}_k < \bar{b}'_m + \bar{\delta}, \\ 0, & \text{otherwise}, \end{cases}$$
$$k = 0 \text{ to } N - 1, \qquad m = 0 \text{ to } N_{mod} - 1. \qquad (14)$$

The weights ν_m serve to normalize each filter by the number of DFT samples it spans:

$$\nu_m = \frac{1}{\left|\left\{k \in \mathbb{Z} : -\bar{\delta} \le \bar{f}_k - \bar{b}'_m < \bar{\delta}\right\}\right|}, \quad m = 0 \text{ to } N_{mod} - 1, \qquad (15)$$

where $|\cdot|$ indicates set cardinality in this context.

Using (14) we can show that the upper slope of filter $\Phi_{\cdot,i}$ and the lower slope of filter $\Phi_{\cdot,i+1}$ intersect at

$$\frac{\nu_i(\bar{b}'_i + \bar{\delta}) + \nu_{i+1}(\bar{b}'_{i+1} - \bar{\delta})}{\nu_i + \nu_{i+1}} \quad \log_2(\text{Hz}). \qquad (16)$$

In the simplified case of $\nu_i = \nu_{i+1}$ (16) becomes the midpoint of $\bar{b}'_i$ and $\bar{b}'_{i+1}$.

We can also use (14) to show that filter centered at $\bar{b}'_i$ has a peak value of ν_i and at the simplified intersection locations $(\bar{b}'_i + \bar{b}'_{i+1})/2$ it has value $\nu_i/\sqrt{2}$. Thus, at the simplified intersection location, the value of the filter relative to the peak value is $1/\sqrt{2}$ or -3.0 dB.

REFERENCES

[1] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Addison-Wesley, 1987.

[2] D. S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 821–831, 2005.