SPHINX III Signal Processing Front End Specification

31 August 1999

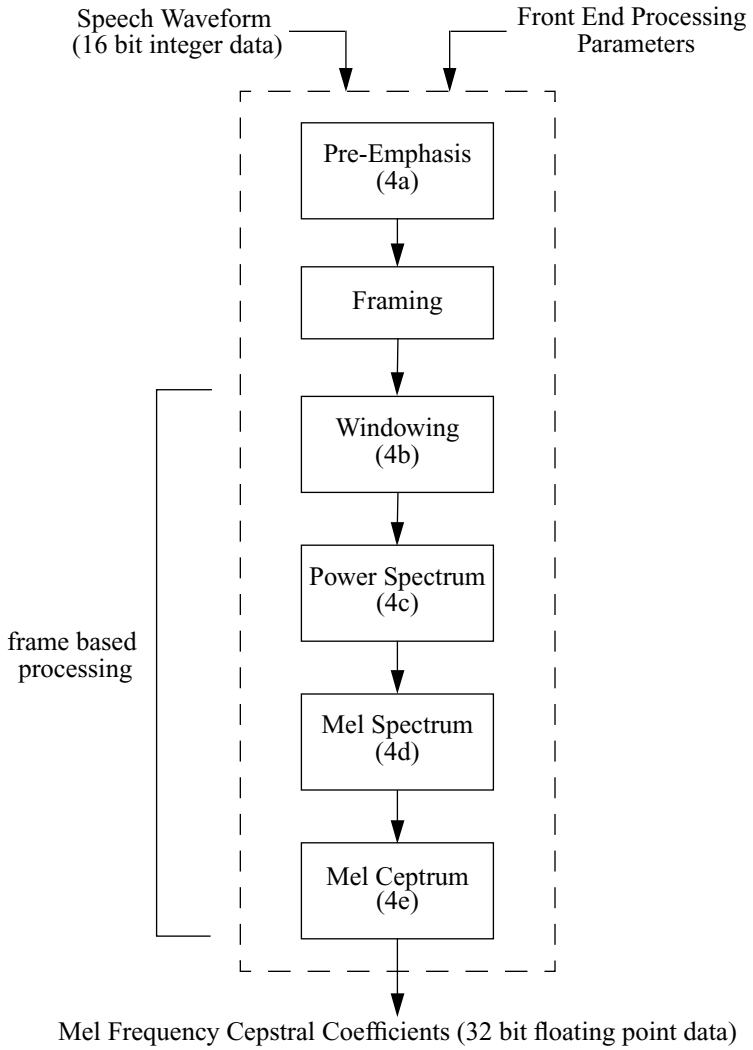
Michael Seltzer (mseltzer@cs.cmu.edu)
CMU Speech Group

1. Introduction

This document describes the signal processing front end of the SPHINX III speech recognition system. The front end transforms a speech waveform into a set of features to be used for recognition, specifically, mel-frequency cepstral coefficients (MFCC).

2. Block Diagram

Below is a block diagram of the feature extraction operations performed by the SPHINX III front end.



3. Front End Processing Parameters

The following parameter structure must be completed by the user prior to using the front end. Any parameter that is set to 0 will be set to its default value (see section 6).

```
typedef struct{
    float32 SAMPLING_RATE;
    int32 FRAME_RATE;
    float32 WINDOW_LENGTH;
    int32 FB_TYPE;
    int32 NUM_CEPSTRA;
    int32 NUM_FILTERS;
    int32 FFT_SIZE;
    float32 LOWER_FILT_FREQ;
    float32 UPPER_FILT_FREQ;
    float32 PRE_EMPHASIS_ALPHA;
} param_t;
```

4. Front End Processing

4a. Pre-Emphasis

The following FIR pre-emphasis filter is applied to the input waveform:

$$y[n] = x[n] - \alpha x[n-1]$$

α is provided by the user or set to the default value. If $\alpha = 0$, then this step is skipped. In addition, the appropriate sample of the input is stored as a history value for use during the next round of processing.

The remaining operations are done on a frame basis.

4b. Windowing

The frame is multiplied by the following Hamming window:

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$$

N is the length of the frame.

4c. Power Spectrum

The power spectrum of the frame is computed by performing a DFT of length specified by the user, and then computing its magnitude squared.

$$S[k] = (\text{real}(X[k]))^2 + (\text{imag}(X[k]))^2$$

4d. Mel Spectrum

The mel spectrum of the power spectrum is computed by multiplying the power spectrum by each of the of the triangular mel weighting filters (see section 5) and integrating the result.

$$\tilde{S}[l] = \sum_{k=0}^{N/2} S[k]M_l[k] \qquad l = 0, 1,...,L\text{-}1$$

N is the length of the DFT, and L is total number of triangular mel weighting filters.

4e. Mel Cepstrum

A DCT is applied to the natural logarithm of the mel spectrum to obtain the mel cepstrum:

$$c[n] = \sum_{i=0}^{L-1} \ln(\tilde{S}[i]) \cos\left(\frac{\pi n}{2L}(2i+1)\right) \quad c = 0,1,...,C\text{-}1$$

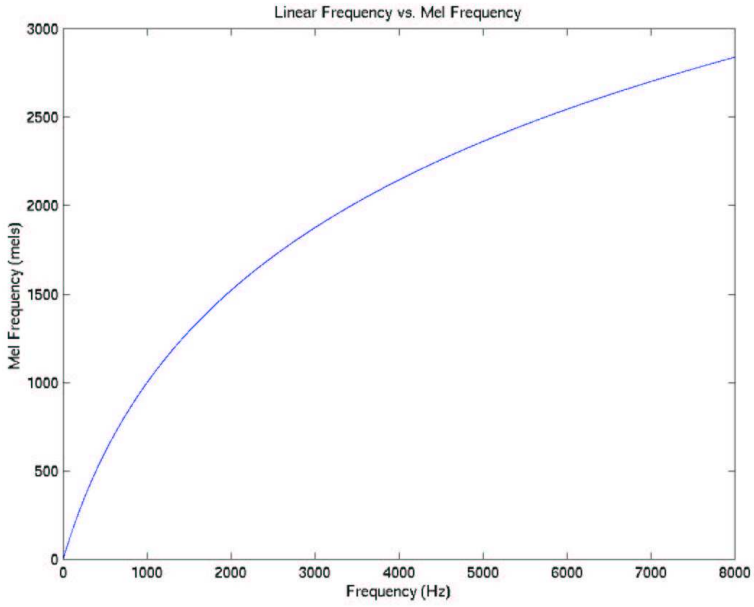
C is the number of cepstral coefficients.

5. Defining the Mel Filterbank

The mel scale filterbank is a series of L triangular bandpass filters that have been designed to simulate the bandpass filtering believed to occur in the auditory system. This corresponds to series of bandpass filters with constant bandwidth and spacing on a mel frequency scale. On a linear frequency scale, this filter spacing is appoximately linear up to 1kHz and logarithmic at higher frequencies. The following warping function tranforms linear frequencies to mel frequencies:

$$mel(f) = 2595 \log\left(1 + \frac{f}{700}\right)$$

A plot of the warping function is shown below.



A series of *L* triangular filters with 50% overlap are constructed such that they are equally spaced on the mel scale spanning [$mel(f_{min})$, $mel(f_{max})$] where f_{min} and f_{max} are set by the user or the default values. The triangles are all normalized so that they have unit area.

6. Signal Processing Front End Default Values

These are the default values for the current SPHINX III front end:

Parameter	Default Value
Sampling Rate	16000.0 Hz
Frame Rate	100 Frames/sec
Window Length	0.025625 sec
Filterbank Type	Mel Filterbank
Number of Cepstra	13
Number of Mel Filters	40
DFT Size	512
Lower Filter Frequency	133.33334 Hz
Upper Filter Frequency	6855.4976 Hz
Pre-Emphasis α	0.97

7. References

D. O'Shaughnessy. Speech Communication - Human and Machine. Addison-Wesley, Reading, 1987.

L. Rabiner, B. Juang. Fundamentals of Speech Recognition. Prentice Hall, New Jersey, 1993

A. Oppenhein, R. Schaefer, J. Buck. Discrete-Time Signal Processing. Prentice Hall, New Jersey, 1999