ELIGIBILITY PROPAGATION IN MULTI-LAYER RECURRENT SPIKING NEURAL NETWORKS

WERNER VAN DER VEEN

Faculty of Science and Engineering
University of Groningen

June 2021 – classicthesis v4.6

Werner van der Veen: *Eligibility Propagation in multi-layer recurrent spiking neural networks,* Master's Thesis, © June 2021

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

1 METHOD

1.1 DATA PREPROCESSING

1.1.1 *The TIMIT speech corpus*

TIMIT is a speech corpus that contains phonemically transcribed speech (Garofolo et al., 1993). It contains 6300 sentences, 10 spoken by each of the 630 speakers. The male and female speakers lived in 8 different geographical regions in the United States during their childhood years, see Table 1.1.

DIALECT REGION	#MALE	#FEMALE	TOTAL
1 (New England)	31 (63%)	18 (27%)	49 (8%)
2 (Northern)	71 (70%)	31 (30%)	102 (16%)
3 (North Midland)	79 (67%)	23 (23%)	102 (16%)
4 (South Midland)	69 (69%)	31 (31%)	100 (16%)
5 (Southern)	62 (63%)	36 (37%)	98 (16%)
6 (New York City)	30 (65%)	16 (35%)	46 (7%)
7 (Western)	74 (74%)	26 (26%)	100 (16%)
8	22 (67%)	11 (33%)	33 (5%)
All	438 (70%)	192 (30%)	630 (100%)

Table 1.1: Distibution of speakers' dialect regions and sexes. Speakers of dialect region 8 moved around a lot during their childhood.

The sentence text can be categorized into 2 *dialect* sentences, 450 *phonetically compact* sentences, and 1890 *phonetically-diverse* sentences.

The dialect sentences, which are spoken by all speakers, are designed to expose the dialectical variants of the speakers. The phonetically compact sentences are designed to include many pairs of phones. The phonetically diverse sentences are taken from the Brown Corpus (Kucera, Kučera, and Francis, 1967) and the Playwrights Dialog (Hultzsch et al., 1964) in order to maximize the number of allophones (i. e., different phones used to pronounce the same phoneme). Table 1.2 lists an overview of the distribution of the number of speakers per sentence type.

Each of the sentences is encoded in as a waveform signal in `.wav`format, and is accompanied by a corresponding text file indicating what phones are pronounced in the waveform, and between which pair of sample points.

SENTENCE TYPE	#SENTENCES	#SPEAKERS	TOTAL
Dialect	2	630	1260
Compact	450	7	3150
Diverse	1890	1	1890
Total	2342		6300

Table 1.2: Distribution of sentence types.

1.1.2 *Engineering features*

In this subsection, we describe the preprocessing pipeline as in Fayek, 2016, which can be summarized by applying a pre-emphasis filter on the waveforms, then slicing the waveform in short frames, taking their short-term power spectra, computing 26 filterbanks, and finally obtain 12 Mel-Frequency Cepstrum Coefficients (MFCCs). We align these MFCCs with the phones found in the TIMIT dataset. An example of a waveform signal is given in Figure **??**.

PRE-EMPHASIS In speech signals, high frequencies generally have smaller magnitudes than lower frequencies. To balance the magnitudes over the range of frequencies in the signal, we apply a pre-emphasis filter $y(t)$ on the waveform signal $x(t)$:

$$y(t) = x(t) - 0.97x(t-1). \tag{1.1}$$

This procedure yields the additional benefit of improving the signal-to-noise ratio.

FRAMING The waveforms, which are sampled at a rate f_s of 16 kHz, cannot be directly used as input to the model, because they are too long—a typical sentence waveform contains in the order of tens of thousands of samples. Furthermore, the samples are not very informative, because they represent the sound wave of the uttered sound. These sounds are filtered by the shape of the vocal tract, which manifests itself in the envelope of the short time power spectrum of the sound. This power spectrum representation describes the power of the frequency components of the signal over a brief interval. We assume the frequency components to be stationary over short intervals—in contrast to the full sentence, which carries its meaning because it is non-stationary. Therefore, we transform the waveform signals into series of frequency coefficients of short-term power spectra. To obtain multiple short-term power spectra over the duration of the waveform, we slice it up into brief overlapping frames.

Every 160 samples (equivalent to 10 ms) of a pre-emphasized signal we take an interval frame of 400 samples (equivalent to 25 ms). This means that the frames overlap by 25 ms. The waveform is zero-padded such that the last frame also has 400 samples. By this process, we obtain

signal frames $x_i(n)$, where n ranges over 1–400, and i ranges over the number of frames in the waveform.

Then, we apply a Hamming window with the form

$$w[n] = a_0 - a_1 \cos\left(\frac{2\pi n}{N-1}\right), \tag{1.2}$$

where N is the window length of 400 samples, $0 \leq n < N$, $a_0 = 0.53836$, and $a_1 = 0.46164$. A plot of this window is given in Figure ??. This

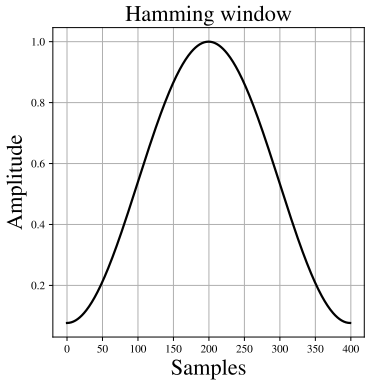

Figure 1.1: The magnitudes of the DFT of a frame.

window is applied to reduce the spectral leakage, which manifests itself though sidelobes in the power spectra. Applying the Hamming window reduces the sidelobes to near-equiripple conditions (Smith, accessed <date>)..

plot for illustration

SHORT-TERM POWER SPECTRA We obtain the power spectra P_i for each frame by first taking the absolute K-point discrete Fourier transform (DFT) of the frame samples $x_i(n)$

$$X_k = \left| \sum_{n=0}^{N-1} x_i(n) \cdot e^{-\frac{i2\pi}{N} kn} \right|, \tag{1.3}$$

where $K = 512$. This yields the magnitudes of the DCT of the frames (an example is illustrated in Figure ??).

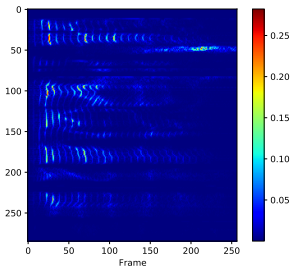

Figure 1.2: The magnitudes of the DFT of a frame.

We obtain the power spectrum using the equation

$$P = \frac{{X_k}^2}{K}, \tag{1.4}$$

an example of which is shown in Figure ??.

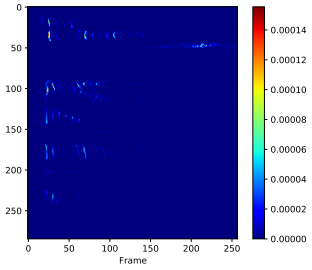

Figure 1.3: A power spectrum of a frame.

MEL FILTERBANK We then transform the short-term power spectra to Mel-spaced filterbanks. The Mel scale is a scale of pitches that are perceptually equal in distance (Stevens, Volkmann, and Newman, 1937). This is in contrast to the frequency measurement, in which the human cochlea can better distinguish lower frequencies better than higher ones. The aim of converting to the Mel scale is to make every filterbank coefficient feature equally informative, thereby improving the learning performance of the model.

The Mel-spaced filterbank is a set of 40 triangular filters that we apply to each frame in P.

To compute the Mel-spaced filterbank we choose lower and upper band edges of 0 Hz and $f_s/2 = 8$ kHz, respectively, and convert these to Mels using

$$m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \tag{1.5}$$

where f is the frequency in Hz. We obtain a lower band edge of 0 Mels and an upper band edge of approximately 2835 Mels.

We begin obtaining the 40 filterbanks by spacing 42 points $\mathbf{m}$ linearly between these bounds (inclusive). Hence, we obtain 42 points spaced exclusively between the bounds.

Then, we convert each point m back to Hz using

$$f = 700\left(10^{m/2595} - 1\right). \tag{1.6}$$

We round each resulting Mel-spaced frequency f to their nearest Fourier transform bin b using

$$b = \lfloor (K+1)\mathbf{f}/fs \rfloor \tag{1.7}$$

The resulting 40 filterbanks with their corresponding Mels and frequencies are listed in Table A.1.

The i^{th} filter in filterbank H_i is a triangular filter that has its lower boundary at b_i Hz, its peak at b_{i+1} Hz, and its upper boundary at b_{i+2} Hz. For other frequencies, they are 0. Therefore, the filterbank can be described by

$$H_i(k) = \begin{cases} 0 & k < b_i \\ \frac{k-b_i}{b_{i+1}-b_i} & b_i \leq k < b_{i+1} \\ 1 & k = b_{i+1} \\ \frac{b_{i+2}-k}{b_{i+2}-b_{i+1}} & b_{i+1} < k \leq b_{i+2} \\ 0 & b_{i+2} < k \end{cases}, \tag{1.8}$$

where $0 \leq k \leq \frac{K}{2}$. These Mel-spaced filters is shown in Figure ??.

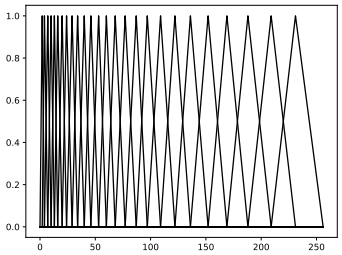

Figure 1.4: The Mel-spaced filterbanks.

We obtain a spectrogram S of the frame (see e. g.Figure ??) after applying the filterbank to the short-term power spectrum.

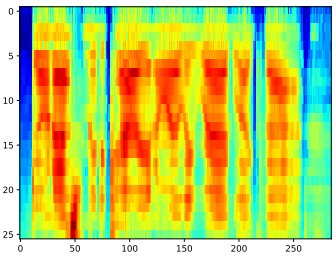

Figure 1.5: An example of a spectrogram.

MEL-FREQUENCY CEPSTRAL COEFFICIENTS We observe that the coefficients in the spectrograms are strongly correlated, which would negatively impact the learning performance of the model .

why?

Therefore, we apply the DCT again to decorrelate the coefficients and obtain the power cepstrum C of the speech frame:

$$C(n) = \left| \sum_{n=0}^{N-1} S(n) \cdot e^{-\frac{i2\pi}{N} kn} \right|. \tag{1.9}$$

We discard the first coefficient in C, because it is the average power of the input signal and therefore carries little meaning. We also discard coefficients higher than 13, because they represent only fast changes in the spectrogram and increase the complexity of the input signal while adding increasingly less meaning to it.An example of the remaining MFCC components is shown in Figure ??.

source?

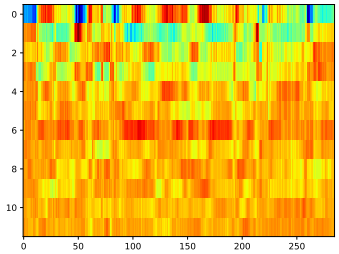

Figure 1.6: An example of Mel-frequency cepstral coefficients that are given as input to the system.

Finally, we balance the final MFCCs by centering them around 0 per frame. An example of the final MFCCs is given in ??.

TARGET OUTPUT The target output of the model is a frame-wise representation of the phones that are uttered in a sentence. The TIMIT corpus contains text files indicating in what order phones occur in a sentence, and their starting and ending sample points.

These phones are discretized into frames such that they align correctly with the MFCCs. They are represented in one-hot vector encoding. Since the dataset contains 61 different phones, this is also the length of these vectors.

Figure ?? illustrates the waveform data and its framewise aligned MFCCs and target output.

side-by-side with original text and phonemes, label as fig:source_mfcc_target

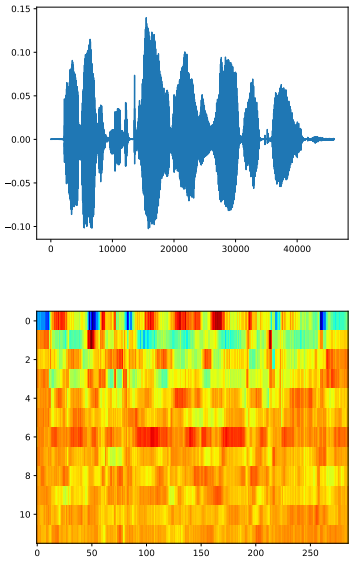

Figure 1.7: An alignment of a sample signal with its MFCCs and target phones.

APPENDIX A

MELS	HZ	FILTERBANK
0	0	0
105	68.5	2
210	143.7	4
315	226.2	7
420	316.8	10
525	416.3	13
630	525.5	16
735	645.4	20
840	777	24
945	921.5	29
1050	1080.1	34
1155	1254.4	40
1260	1445.4	46
1365	1655.3	53
1470	1885.7	60
1575	2138.6	68
1680	2416.3	77
1785	2721.2	87
1890	3055.9	97
1995	3423.3	109
2100	3826.7	122
2205	4269.5	136
2310	4755.7	152
2415	5289.4	169
2520	5875.3	188
2625	6518.6	209
2730	7224.8	231
2835	8000	256

Table A.1: Conversion table between linearly spaced Mels and their corresponding frequencies and filterbank boundaries.

BIBLIOGRAPHY

Fayek, Haytham M. (2016). *Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between*. URL: https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html.

Garofolo, John S et al. (1993). "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1." In: *STIN* 93, p. 27403.

Hultzsch, Eugen et al. (1964). *Tables of transitional frequencies of English phonemes*. Urbana: University of Illinois Press.

Kucera, Henry, Henry Kučera, and Winthrop Nelson Francis (1967). *Computational analysis of present-day American English*. Brown university press.

Smith, Julius O. (accessed <date>). *Spectral Audio Signal Processing*. online book, 2011 edition. http://ccrma.stanford.edu/~jos/sasp/.

Stevens, Stanley Smith, John Volkmann, and Edwin B Newman (1937). "A scale for the measurement of the psychological magnitude pitch." In: *The Journal of the Acoustical Society of America* 8.3, pp. 185–190.