

Learning the Mel-scale and Optimal VTN Mapping

Terri Kamm, Hynek Hermansky
and Andreas G. Andreou

Abstract - Traditional front-end features for speech recognition employ Mel-warped cepstrum. In this experiment we ask: Why Mel-warped? Using carefully controlled vowel database, we design an experiment that involves nearly one million monotonic warps and classifiers and observe the class of warps which gives the best classification to the vowel recognition task. The Mel-warp is indeed a member of this class of discriminative warps.

1. Introduction

Mel-scale grouping is a commonly employed dimensionality reduction technique in automatic speech recognition systems. The grouping follows a filtering operation that converts the time series representation of speech into a frequency representation. Beginning with a large number-typically 128- inputs, linearly spaced in the frequency scale, one obtains a smaller set of outputs -typically 14-that are non-linearly spaced on the frequency axis and yet capture the essential information for speech recognition.

2. Background

To understand how the warping was incrementally modified in this experiment, the reader must first understand in general how cepstral coefficients and mel-warped cepstral coefficients (MFCCs) are calculated for use in typical speech recognition systems. (A very good reference is [Deller 1993].) This paper will specifically discuss how the mel-warped cepstrum is calculated in the Hidden-Markov Toolkit (HTK) system [Young 1995].

Given a frame of speech, the following steps compute the cepstral coefficients:

- 1) Window the speech frame, using a Hamming (or other) window.
- 2) Do zero padding to achieve a frame length suitable for an FFT.
- 3) Do the FFT.
- 4) Find the Power Spectrum.
- 5) Take the Log of the Power Spectrum. (This is the cepstrum.)
- 6) Inverse FFT. (The results are the cepstral coefficients.)

The "Mel" is a unit of measure of perceived pitch or frequency of a tone. The Mel scale was developed by Stevens and Volkman (1940) as a result of a study of human auditory perception (the field of psychoacoustics) using the following procedure:

- 1) Choose the reference frequency as 1000 Hz and designate it "1000 Mels".
- 2) Listeners were then presented a signal and asked to change it's frequency until the pitch they perceived was twice the reference, then 10 times the reference, etc. and then half the reference, 1/10 the reference, etc.
- 3) From this data, the mel scale was constructed.

The mel scale is approximately linear below 1 kHz and logarithmic above. This is approximated in HTK using the following formula, and shown in Figure 1:

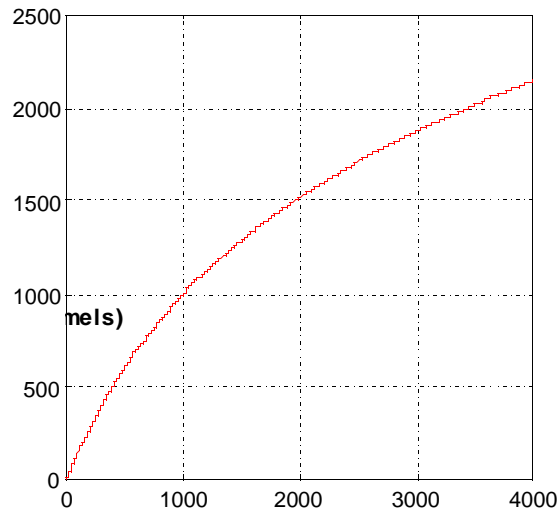


Figure 1: The Mel Scale

$$\text{Mel}(f) = 1127 \log\left(1 + \frac{f}{700}\right)$$

To apply this information to the 6 steps which compute the cepstral coefficients, suppose that it is desired to compute the cepstral coefficients at frequencies distributed linearly on the range 0-1000 Hz, and logarithmically above 1000 Hz. To do this, we could "oversample" the frequency axis by using an FFT in step 2 which is 2 to 4 times larger than the speech frame requires, and select the frequency components which represent the said distribution. The remaining components could then be set to zero (after step 4, in practice set to 1 because of the log), and the remaining steps could be performed.

Rather than throwing out information with zeroing, it is desirable to once again mimic the human auditory system by applying a second psychoacoustic principle. It has been found that the perception of a particular frequency, say f , by the auditory system is influenced by energy in a critical band of frequencies around f [Schroeder 1977]. Further, the bandwidth of a critical band varies with frequency, beginning at about 100 Hz for frequencies below 1000 Hz and then increasing logarithmically above 1000 Hz. This implies that a sensible thing to do is to use the log total energy in critical bands around the mel frequencies as input to step 6, the inverse FFT.

In HTK, this is realized by using a Fourier transform based filterbank, which is derived from the mel scale. A 20 channel filterbank for an 8000 Hz signal is shown in Figure 2. To implement this filterbank, the window of speech is transformed using an FFT and the magnitude is taken. These magnitudes are then "binned" by correlating them with each triangular filter. Here "binning" means that each FFT magnitude coefficient is multiplied by the corresponding filter gain and the results accumulated. Thus, each bin holds a weighted sum representing the spectral magnitude in that filterbank channel.

Finally, the mel-weighted cepstral coefficients are calculated from the log filterbank amplitudes, denoted m_j , using the Discrete Cosine Transform

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^N m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right), \quad i = 1..number \text{ of cepstral coefficients}$$

where N is the number of filterbank channels.

Note that in this experiment, we used the mel-weighted log power spectrum, rather than the mel-weighted cepstral coefficients.

3. Experimental Framework

3.1 Database

The Western Michigan vowel database [Hillenbrand 1995], from now on denoted WMVD, was used in the experiment described in this paper. This database was collected to replicate the [Peterson 1952] vowel study. The database consists of 45 men, 48 women, and 46 ten- to 12-year-old children, speaking 12 vowels in the /hVd/ context. The acoustic signal, sampled at 16000 Hz, as well as hand measurements of formants, vowel duration and vowel steady-state point make up this database.

For this experiment we selected the adult data and for each speaker calculated a

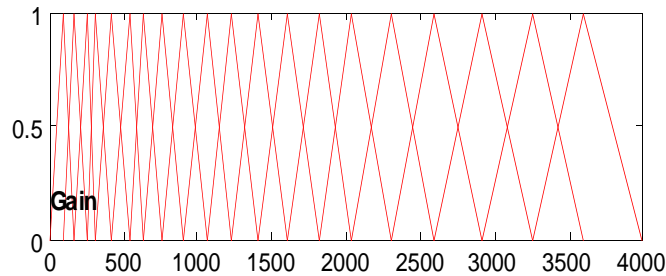


Figure 2: Mel Scale Filter Bank

256-point power spectrum centered at the steady-state point of each vowel, having first applied a Hamming window. Then we threw away the upper 128-points, leaving a linear power spectrum in the range of 0-4000 Hz, which gave us a feature vector which is most similar to the features used in recognition of telephone speech.

3.2 Monotonic Warping Vectors

Given a 128-point power spectrum, we wanted to try all possible warpings which would yield 16 channels. This is certainly a large number, much more than we could compute within a reasonable time frame. In order to reduce the number of possible warpings, we narrowed our focus to monotonic warpings only.

Given the size of the spectrum, $N(=128)$, and the number of desired warped channels, $M(=16)$, a valid warping vector $W_j = [C_1 C_2 C_3 \dots C_M]$, where each element represents the size of the channel, will obey the following rules:

$$C_i \leq C_{i+1}, i = 1..(M - 1)$$

$$\sum_{i=1}^M C_i = N$$

So an algorithm to compute the warping vectors goes as follows:

- 1) Initialize C_i to 1, for $i=1..M-1$. Then choose C_M such that all C_i 's sum to N .
- 2) Increment, with carry, the C_i for $i=1..M-1$, always adjusting C_M such that all C_i 's sum to N , and verifying that the monotonic rule is obeyed.
- 3) In the event of a carry at place i , reinitialize C_i to C_{i-1} .

The following example has 5 warping vectors:

$$N = 8, M = 5$$

$$W_1 = 1 \quad 1 \quad 1 \quad 5$$

$$W_2 = 1 \quad 1 \quad 2 \quad 4$$

$$W_3 = 1 \quad 1 \quad 3 \quad 3$$

$$W_4 = 1 \quad 2 \quad 2 \quad 3$$

$$W_5 = 2 \quad 2 \quad 2 \quad 2$$

Additional tradeoffs were made to achieve a reasonable number of warping vectors. First, the initial vector was set to [1 1 1 1 1 1 1 2 2 2 2 3 3 3 93]. For the first 15 channels, these numbers also represent the increment for that channel. Second, the lowest 10 spectral bins were removed from consideration, thereby making $N = 118$. This

is justified by the fact that telephone speech usually does not contain speech information in the lower 300 Hz. This resulted in 845,000 warping vectors.

3.3 Maximum-Likelihood Classifier

Discrimination between the twelve vowel classes was done using a maximum-likelihood classifier. The means and covariances for each class were estimated, assuming identical covariance across classes.

4. Experiments

4.1 Why Mel?

From the background information, we saw that the development of the Mel-scale was based on human auditory experiments. This doesn't necessarily imply that the Mel-scale should also be good for machine recognition of speech. In order to investigate what scalings make sense for machine recognition, we devised the following experiment:

- 1) Divide the speakers into train and test sets.
- 2) For each warping vector:
 - a. Compute new features by summing over the bins as specified by the warping vector and taking the log of the result.
 - b. Train a model of each vowel class using the training set.
 - c. Determine the error rate on the test set.

The top 32 warpings with the lowest error rates are shown in Figure 3. The Mel scale is included in this class of best warping and is highlighted in the figure. There was no evidence to suggest that the Mel-scale is an inappropriate choice for automatic discrimination of vowels.

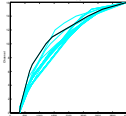


Figure 3: Top 32 Warping, Mel highlighted

4.2 Why Linear?

In [Cohen 1994, Cohen 1995], it was reported that a linear normalization in the frequency domain could be successfully used to compensate somewhat for the first-order effects of differences in vocal tract length. Now we pose the question, "Is a linear normalization the optimal choice?"

We investigated the question with the following experiment:

- 1) Choose the male speakers for training and the female for the test set.
- 2) Using a mel-scaled feature, train a model of each vowel class using the training set.
- 3) For each warping vector:
 - a. Compute new test features by summing over the bins as specified by the warping vector and taking the log of the result.
 - b. Determine the error rate on the test set.

Our experiments suggest that the linear warping is sub-optimal and does not necessarily yield the best frequency grouping.

5. References

- Cohen, J, Kamm, T, Andreou, A, (1994). "An experiment in systematic speaker variability", *Final Day Rev., DOD Speech Workshop on Robust Speech Recognition*, 1994.
- Cohen, J, Kamm, T, Andreou, A, (1995). "Vocal Tract Normalization in speech recognition: Compensating for systematic speaker variability". *J. Acoust. Soc. Am.*, **97** (5), Pt. 2, pp. 3246-3247.
- Deller, JR, Proakis, JG, Hansen, JHL, (1993). *Discrete-Time Processing of Speech Signals* (Macmillan Publishing Company, New York, NY). Chapter 6: Cepstral Analysis.
- Hillenbrand, J, Getty, LA, Clark, MJ, Wheeler, K, (1995). "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.*, **97** (5), Pt. 1, pp. 3099-3111.
- Peterson, GE, Barney, H, (1952). "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.* **24**. pp. 175-184.
- Schroeder, MR, (1977). "Recognition of complex acoustic signals," *Life Science Research Reports*, Vol. 55, pp. 323-328
- Stevens, SS, Volkman, J, (1940). "The relation of pitch to frequency," *American Journal of Psychology*, Vol. 53, pg. 329.
- Young, S, et. al. (1995). *The HTK Book for HTK V2.0* (Published by Entropic. Provided with purchase of software.)