

# An approach of Feature Techniques using Warped frequencies

Saimir Tola<sup>1)</sup>, Alfred Daci<sup>1)</sup>

Polytechnic University of Tirana  
Faculty of Mathematical and Physical Engineering,  
Department of Mathematical Engineering, Tirane,  
saimir\_tola@yahoo.com

**Abstract:** In this paper we will use different feature extraction approach in order to compare the stability of the Automatic Speech Recognition (ASR) on different SNR noise, in different situations. By using the classic feature extraction like: Mel filter Bank, Fast Fourier Transformation (FFT), Discrete Fourier Transformation (DFT). After we implement this methods for the different situations then we will warp the frequency of this methods directly on the spectrum in order to see if we get better results. The results will be compared with one another using the Word Error Rate algorithm (WER).

**Keywords:** Automatic Speech Recognition, Mel Filter Bank, Fast Fourier Transform

## 1. Introduction

Information processing machines have become normal talking nowadays. Though, the current styles of human machine communication are focused more towards living with the limitations of computer input/output devices rather than the ease of humans. Speech is the main mode of communication among people. On the other hand, prevalent means of input to computers is through a keyboard or a mouse. If computers could listen to human speech and carry out their commands. Automatic Speech Recognition is the process of deriving the transcription of an utterance, given the speech waveform. Speech understanding goes further, and collects the meaning of the word in order to carry out the speaker's command. Presence of background noise reduces signal to noise ratio. Background speech of neighbors rise to significant confusions among speech sounds. Speech recorded by a desktop speakerphone not only captures speaker's voice but also multiple echo from walls and other reflecting surfaces [1]

### 1.1 Mel filter Bank

Speech processing shows an significant role in any speech system whether it's Automatic Speech Recognition or speaker recognition or something else. Mel-Frequency Cepstral Coefficients were very common features for a long time; but more recently, filter banks are becoming increasingly popular. In this post, I will discuss filter banks and MFCCs and why are filter banks becoming increasingly popular. Calculating filter banks and MFCCs involve somewhat the same procedure, where in both cases filter banks are computed and with a few more extra steps MFCCs can be obtained. In a nutshell, a signal goes through a pre-emphasis filter; then gets sliced into (overlapping) frames and a window function is applied to each frame; afterwards, we do a Fourier transform on each frame and calculate the power spectrum; and subsequently compute the filter banks. To obtain MFCCs, a Discrete Cosine Transform is applied to the filter banks retaining a number of the resulting coefficients while the rest are discarded.

### 1.2 Discrete Fourier Transformation

Discrete Fourier Transform is one of the most important and dependable mathematical workers in signal processing. DFT of a signal is defined as DFT is used to sample the spectrum at a range of frequencies. Inappropriately, the spectrum is oversampled at a finer resolution and every output of the filter bank which is a power spectral magnitude is processed as a weighted sum of its adjacent values. DFT uses averaging to tool a spectral smoothing function. Averaging is one technique widely used for spectral smoothing. The technique of averaging is often utilized in the Mel scale frequency domain if a DFT is used because the added computational load is minimal.

By performing averaging in the log domain or log power values as opposed to spectral amplitudes, is beneficial for spectral analysis as shown in figure 1.

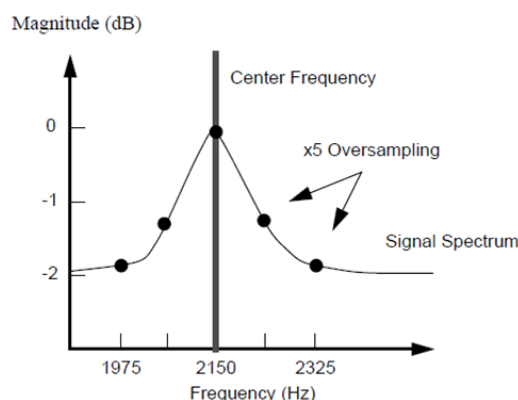


Figure 1: Log domain averaging impact on Spectral Analysis [2]

The combination of pure tones or sine waves results in a complexity of sound. A frequency analysis of such a sound often attempts to identify the original pure tones. The Fourier Transform was designed by the French mathematician Fourier in the 1820's and remains the primary method for carrying out frequency analyses of sounds as well as other phenomena. A number of different ways of performing the Fourier Transform have been developed including the Discrete Fourier Transform and the Fast Fourier Transform and Sparse Fourier Transform. These are designed for working with digital signals such as speech signals.

### 1.3 Warping the spectrum directly

It is stated that warping the DFT directly instead of using filter bank averaging provides a more precise estimate of the perceptual scales. This was a study on additive noise degradation in ASR systems. There is a large body of research on improving the robustness of speech recognition systems under adverse acoustic environments. Environment compensation methods can be applied at the front end feature domain or at the back end model domain or can be applied on areas of the ASR system. Fast Fourier Transform and Discrete Fourier Transform are basically alike; with the only difference being that FFT is faster. Warped DFT or FFT based features have been found to provide lower recognition error rates than the DFT based cepstral features. In the conventional MFCC front-end, processing of a speech signal begins with the pre-processing stage. This involves DC removal and pre-emphasis using a first-order high-pass filter with a transfer function followed by a Fourier transform being applied as was previously discussed. Transforming a linear frequency scale to a non-linear frequency scale is called

frequency warping. One technique to achieve frequency warping is to apply a nonlinearly-scaled filter bank, such as a Mel filter bank, to the linear frequency representation. Another way is to use a conformal mapping, such as the bilinear transformation which preserves the unit circle [3].

$$H(z) = \frac{z^{-1} - \alpha'}{1 - \alpha' z^{-1}}, \quad \forall -1 < \alpha' < 1$$

The equation above is an example of warped DFT. DFT is achieved by applying the FFT algorithm. In warped DFT or FFT the positions of the frequency peaks are modified by using an all-pass transformation to warp the frequency axis. Then, uniformly-spaced points on the warped frequency axis are similar to non-uniformly spaced peaks on the original frequency axis. By picking the warping parameters sensibly, one can place some of the frequency samples in close proximity to each other to provide higher resolution in the frequency range of interest without increasing the length of the DFT. Utilizing this frequency warping, one can improve the spectral representation of speech signals in the low frequency region [4].

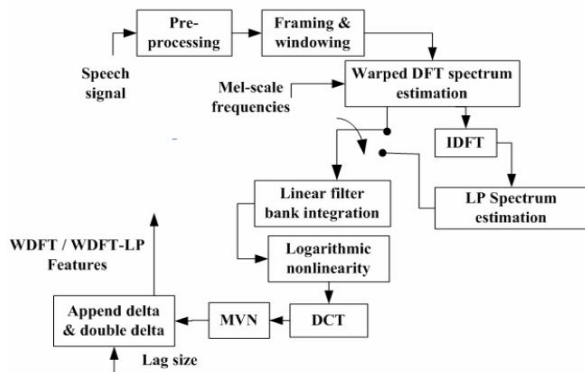


Figure 2: Extraction of warped DFT-based cepstral features

Warping the DFT spectrum directly without using filter bank averaging provides a more precise approximation of the perceptual scales [5]. Since the spectrum is already pre-warped using Mel-frequency warping, the nonlinearly-spaced triangular-shaped Mel-frequency filter bank is replaced by a filter bank of uniformly spaced, half-overlapping triangular filters, to provide spectral smoothing. The figure below shows running speech spectra of (a) clean and (b) noisy speech signals that have been corrupted by babble noise with a signal-to-noise ratio of 6 dB, obtained using DFT, WDFT, and WDFT-LP spectrum estimators. Based on this visual examination, WDFT and WDFT-LP provide more robust spectral estimates compared to DFT and LP methods. Due to reduced degrees of freedom in all-pole modeling [6].

## 2. Experimental Results

To evaluate the robustness and performance of the proposed method in one case of clean and noisy conditions, further to compare with the multiple feature extraction methods, such as Mel filter bank cepstral coefficients, DFT filter bank, FFT, the AURORA database is chosen for the experiments. The speech model is trained by different SNR degrees 0-15 dB of mixed dataset comparing clean and noisy data. To generate the noisy data, the clean dataset is further mixed with three types of noise data, including bus station, public parking garage, and movie theatre. Totally, 7,895 utterances are yielded for training purposes and split equally into 20 subsets, each SNR subset contains 233 utterances

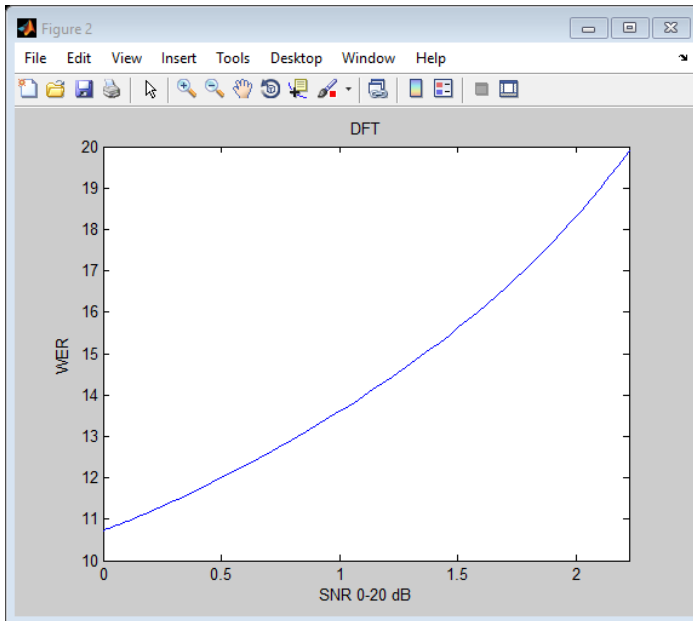
and the sampling rate is 8 kHz, and each subset including one clean data with five SNR types of noisy data, that is, 5 dB, 10 dB, 15 dB, and 20 dB, respectively. In the test part, three types of noisy data with different SNRs on -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB are also built. Each SNR subset contains 9900 utterances and totally 217 600 utterances are generated for testing. For testing the data above us will use the word error rate. We have created these algorithm in Matlab as below, and after the execution we get the data in table 1 and 2..

**Table 1: Mechanical properties of selected powder materials.** *Таблица 1: Източникът на преработката не е نامерен. \*Note: Values in brackets are valid for heat-treated material.*

WER	SNR (dB)					Mean
	0 dB	5 dB	10 dB	15 dB	20 dB	
DFT	1 0.7 2	13.61	1 8.32	2 1.44	23. 57	62 .33
MFC	1 2.1 2	14.97	1 9.02	2 3.21	23. 57	64 .72
WDF	7	10.44	1	1	17.	63
T	.71		4.21	5.22	34	.31

Now we interpolate the data in order to find WER for different SNR data.

```
clear all
clc
format long
syms
x=input('x= ');
y=input('y= ');
xs=input('xs= ');
h=x(end)-x(end-1);x0=x(1);sh=(xs-x0)/h;
k=length(x)-1; b=zeros(k+1,k); yy=y; z=sym(zeros(k,1));
if length(x)~=length(y)
    disp('lengths must be same')
    break
end
for i=1:k
    for j=1:k-i+1
        yy(j)=yy(j+1)-yy(j);
        b(j,i)=yy(j);
    end
end
for i=1:k
    ss=1;
    for j=0:i-1
        ss=ss*(s-j);
    end
    z(i,1)=ss;
end
F=0;
for i=1:k
    F=F+(z(i,1)/factorial(i))*b(1,i);
end
F0=char(F);pn=inline(F0,'s');
disp('      x      y      Dy(forward)s:')
disp('*****')
disp('*****')
disp(['x' y' b])
disp('The NFD interpolation is p(xs=x0+sh)=')
pretty(F+y(1))
disp(' ')
disp(['Newton-forward      interpolation      p(x)      for
x=',{xs},'is',{pn(sh)+y(1)}])
```



### References

- [1] Automatic Speech Recognition Samudravijaya K Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005
- [2] Piccone, J. (1992) Signal Modelling Techniques in Speech Recognition. Texas Instruments  
Shrawankar, U., Thakare, V. (2013) Techniques for feature extraction in speech recognition system: a comparative study.
- [3] Bilmes, J. (2004). "What HMMs can't do," in Proceedings of the Institute of Electronics, Information and Communication Engineers Beyond HMM Workshop, Kyoto, Japan, Vol. 104, SP2004-81-95.
- [4] Çetin, O., Kantor, A., King, S., Bartels, C., Magimai-Doss, M., Frankel, J., and Livescu, K. (2007). "An articulatory feature-based tandem approach and factored observation modeling," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2007)
- [5] T. Wesker, B. Meyer, K. Wagener, J. Anemüller, A. Mertins, and B. Kollmeier. Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines. In Proc. of Interspeech, pages 1273–1276, Lisboa, Portugal, September 2005.
- [6] G. Zhou, M.E. Deisher, and S. Sharma. Causal analysis of speech recognition failure in adverse environments. In Proc. of ICASSP, volume 4, pages 3816–1819, Orlando, Florida, May 2002.