Evaluation Of Mel And Gammatone Filter Banks Used For Spectral Analysis In Comparison With The Direct Use Of Fft

Saimir TOLA, Alfred DACI, Gentian ZAVALANI

Abstract: ***This paper analyses the development of Automatic Speech Recognition systems in relation to the varied types of spectral analysis methods used. A critical evaluation of Mel and Gammatone filter banks used for spectral analysis in comparison with the direct use of FFT spectral values is considered. Research was based on understanding the effectiveness of existing Automatic Speech Recognition systems are specifically focused on Mel and Gammatone filter banks in comparison with FFT spectral values.***

Keywords: Automatic Speech Recognition (ASR), Mel, Gammatone, Fast Furier Transformation (FFT)

1. Introduction

In a speech recognition process, parameterizing the analogue signal is the beginning. Methods or algorithms of speech recognition that develop a coherent representation of the parametric speech signals have evolved and grown over time.

The algorithms that are used today have parameters which mimic the behaviour of natural human and auditory processes and are developed with an aim to maximize recognition performance. Early speech recognition was reliant on speaker dependent technology.

Research aims have transitioned from speaker dependent technology to speaker independent technology over time.

However, the parameterizations used for speaker independent technology were based on techniques used for speaker dependent technology in the past and therefore continued to be utilized even after the transition. For the purposes of parameterization, the key difference used to distinguish the needs of the two forms of speech recognition was that for speaker independent speech, the premium was placed on designing descriptions that are invariant regardless of variations in the speech creator. Therefore, representative parameters of the speaker's voice are less appealing than principle spectral energies of the sound [1].

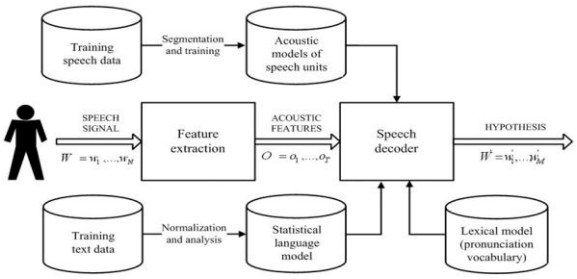


Fig.1: Architecture of a state-of-the-art automatic speech recognition system and its components [2]

Two models are available in an up-to-date ASR scheme: Model training and voice decoding. The objective of system training is to develop and enhance voice acoustics models. For speaker-indépendant ASR, the language requiring a corpus of text or sentence grammar is needed; and a recognition lexicon, which includes a list of identifiable tokens with single or multiple phonetic transcriptions. Acoustic modelling allows the execution of the specific classes of the audio signal of the essential speech units that are context independent, such as monophones, syllables or contexts such as allophones, triphones and pentaphones. Vectors of speech signal features (e.g., melfrequency cepstral coefficients (MFCC), linear prediction coefficients (LPC), perceptual linear prediction coefficients (PLP), bottleneck features (ML), etc.) are extracted from the acoustical signal for dimensionality reduction and probabilistic modeling [2].

2. Methodology

The heart of a Speech Recognition System is Pattern Recognition. Speech recognition when considered via Pattern Recognition consists of two approaches; Signal Modelling and Network Processing. In Signal Modelling, sequences of speech samples are assimilated into organizational vectors using probability space to depict events. Network Processing controls the task of finding the most likely

Revised Manuscript Received on July 05, 2019.

Saimir TOLA, Polytechnic University of Tirana, Dept of Mathematical Engineering, Faculty of Mathematical and Physic Engendering.
Alfred DACI, Polytechnic University of Tirana, Dept of Mathematical Engineering, Faculty of Mathematical and Physic Engendering.
Gentian ZAVALANI, Polytechnic University of Tirana, Department of Mathematical Engineering, Faculty of Mathematical and Physic Engendering.

probable sequence to these events, taking into account probable syntactic constraints.

Spectral shaping, spectral analysis, parametric transformation and statistical modelling make up the four subdivisions that comprise the system of Signal Modelling. Of these subdivisions, the first three can be classified as being straightforward with the fourth division of Statistical Modelling is slightly more complex. Statistical Modelling is divided into a Statistical Modelling System and the Speech Recognition System [1].

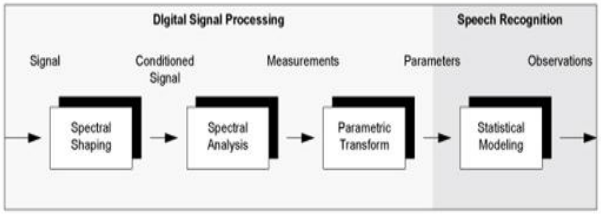


Fig 2: Speech Recognition Process [1]

Three key forces exist in the design and development of Signal Modelling systems. Initially, parameterizations representing salient aspects of the speech signal are looked for. Parameters similar to those used for the human auditory system are preferred. Parameters which imitate the human auditory system are referred to as perceptual parameters. Parameters that are resilient to variations in the channel, speaker, and transducer are subsequently desirable. This nature of resilience is referred to as a issue of stability or invariance.

3. Mel Filter bank

The Mel filter bank is a filter bank in speech recognition systems based on the Mel scale. The Mel filter bank a significant filter bank that is widely used. The frequency/ bandwidths for this filter bank are given in fourth and fifth columns of the figure 3.

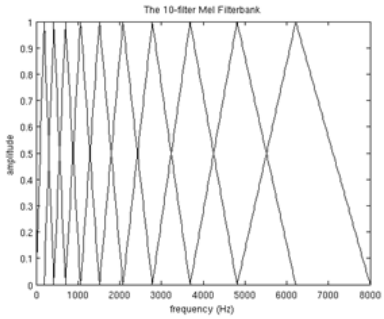


Fig 3: Mel Filter Bank bandwidth [1]

Ten filters of 100 Hz to 1000 Hz are sequentially assigned to the Mel filter (Figure 4). Above 1000 Hz, for each frequency scale replication (octave), there are five filters assigned. The filters are also separated into a log scale. The bandwidths are assigned to ensure that the 3 dB point between the bin and the bin. Only for comparisons and contrast purposes is shown the shaded regions in the table. In general, the first 20 filter bank samples are used only. Each filter is normally installed as a linear phase filter in the Digital Filter Bank, so that the group delay is equal to zero for each filter, and the filter output signal is synchronized in time. For a sequential stage filter execution the filter equations are used. In Figure 5, sound processing by two filters is shown in this type of filter bank. The speech signals with two band pass filter inputs are shown in figure 5. One filter band pass is 250 Hz center and the other is 2500 Hz centered. The image shows obviously, that the output energy varies with the sound type.

Index	Bark Scale		Mel Scale	
	Center Freq. (Hz)	BW (Hz)	Center Freq. (Hz)	BW (Hz)
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	184
13	1850	280	1516	211
14	2150	320	1741	242
15	2500	380	2000	278
16	2900	450	2297	320
17	3400	550	2639	367
18	4000	700	3031	422
19	4800	900	3482	484
20	5800	1100	4000	556
21	7000	1300	4595	639
22	8500	1800	5278	734
23	10500	2500	6063	843
24	13500	3500	6964	969

Fig. 4: A Mel Filter bank with 10 filters [1]

The core advantage of a filter bank is that certain filter outputs can be made compatible with certain classes of speech sounds.

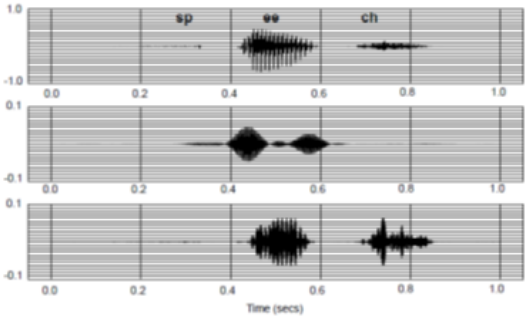


Fig. 5: Fiter bank processing sound signals [1]

Building appropriate feature representations for speech recognition has been an important area of research for many years. The popularity of the Mel filter bank is due to the motivational theories of speech production and speech perception. However, as was stated by Sainath et al [3], although Mel derived features are popular; Mel filter banks are not the correct choice for speech recognition as they are not built for this purpose [4].

The Mel filter bank is designed on the basis of mimicking the auditory and physiological evidence of how humans perceive audio signals. It is therefore, designed from perceptual evidence. In a

statistical modeling framework, the end goal is reducing the word error rate, as it impacts the quality of the ASR system output. In reference to achieving this, filter banks such as Mel and Gammatone which are built on emulating human auditory models are at a disadvantage

4. Fourier Transform Filter Bank

A Fourier transform on the signal is one of the most convenient and effective techniques of computing a non standardized filter bank model of the soundsignal. A transform from Fourier can be used to test the transformed output at the required frequencies.

Discrete Fourier Transform (DFT) is one of the most important and reliable mathematical workers in signal processing. DFT of a signal is defined as

$$S(f) = \sum_{n=0}^{N_s-1} s(n) e^{-j\left(\frac{2\pi f}{f_s}\right)n} \quad (1)$$

DFT is used to sample the spectrum at a range of frequencies. Unfortunately, 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 implement a spectral smoothing function. Averaging is one technique widely used for spectral smoothing [5]. The technique of averaging is often utilized in the Mel scale frequency domain if a DFT is used because the added computational burden 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 6

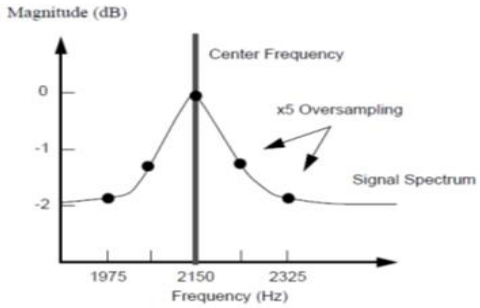


Fig. 6: Log domain averaging impact on Spectral Analysis [1]

Combining pure tones and waves leads to a high level of sound complexity. A sound frequency assessment often tries to define the initial pure tones. Fourier Transform was designed for carrying out frequency analysis of sounds as well as other phenomena by the French mathematician Fourier in the 1820s. There have been created a range of different ways to carry out the Fourier Transform, which include the Discrete Fourier Transformation (DFT). These are intended for digital signals like speech signals.

The Fast Fourier Transform (FFT) algorithm is the most common technique within Fourier transform methods. The FFT was created in the 1960s and the frequency display of a magnitude N signal in the O(N log N) time is calculated. An appropriate way to handle the signal spectrum could be a fast-fourier transform (FFT). The FFT is a computerized deployed Fourier Discrete Transformation, limited to the requirement that the spectrum must be assessed at a discrete frequency set of multiple frequencies:

$$\frac{f_s}{N} \quad (2)$$

These frequencies are referred to as orthogonal frequencies. The principal advantage of the FFT is that it is very fast. It is customary to add an additional processing step. It is hypothesized that areas of maximum vibration or amplitude are given more weight in the human hearing system that are low amplitude areas [1].Therefore, in noisy environments the background noise tends to adversely impact our estimates of the low amplitude areas of the spectrum in a disproportionate manner. In other words, we tend to be more reliant on the reliability of our estimates of the high amplitude areas of the spectrum. Due to this a limit on the dynamic range of the spectrum is usually imposed. This low limit is called the dynamic range threshold. By choosing a specific threshold from the peak in the spectrum, we simply cut or discard estimates below a specific threshold from the highest point in the spectrum[5]. To implement this threshold algorithms are used on the spectrum for methods

based on the Fourier Transform. For the threshholding algorithm, it is necessary that the spectral values are relatively flat prior to deployment. If this step is not taken low energy spectral areas that are productive can be accidentally destroyed. Due to the fact that the spectrum of the audio signal of human speech drops 20 dB each 10 years, having a threshold founded on low frequency energy is realistic [6]. This is due to the fact that when the lowest to highest spectral amplitude range difference is large relevant audio energy at higher frequencies can be destroyed. Therefore FFT deploys a useful technique of setting a threshold and protecting the most productive data for processing sound signal.

5. Experimental Results

In order to assess the robustness and efficiency of the technique in two smooth noisy situations, the AURORA-2 database is selected for experiments in order to compare it to the portable extraction processes Mel filter bank, Gamma tone filter bank, FFT. A combined data set computering clean and noisy is trained with various SNR levels(-5-20 DB Step-5 DB). In order to produce the noisy data, three kinds of noise data, including (1) suburban, (2) vehicle and (3) exhibition hall, are also added to the smooth data set. For practice, the total number of utterances is 9 880, and divided equally in 30 subsets and the sampling rate of each SNR subset is 8kHz, each subset includes 5 dB, 10 dB, 15 dB and 20 dB, including one clean information with five kinds of SNR noisy dataThree kinds of noisy data

with various SNRs are also available in the test section on-5 dB,0 dB, 5 dB, 10 dB, 15 dB and 20 dB. Each SNR sub-set includes 9900 utterances and for testing, there are completely 237 600 utterances. We will use the word error rate for testing the above information. We developed the following algorithm in Matlab and we get information in Tables 1 and 2, following the execution.

```
    function w=WER(h,r)
% this function is for calculation word error rate (WER)
% between the sequence of words h (hypothesis) and the
sequence of words r (references).
% WER determines the performance of an automatic speech
% recognition. The metric Levenshtein is used to calculate the
distance at the level of words.
(H, r) / N where D (h, r) distances Levenshtein between h and
r,
w=[];
switch nargin
 case [{0},{1}]
  warning('There are more than two arguments for ex. of the
function.')
  return
 case [{2}]
  if nargin==2&&ischar(h)
   h=strsplit(h); % realizes the partition of the sentence
  end
  if nargin==2&&ischar(r)
   r=strsplit(r); % realizes the partition of the sentence
  end
  d1=strd(h,r,1);  % distance between h and r case "sensible"
  d2=strd(h,r,2);   % distance between h and r case "not
sensitive"
  d3=numel(r);     % calculates the number of words in r
  if isempty(r)
   warning('Reference should have words.')
  end
  w=[d1,d2]/d3;  % calculation of WER
 otherwise
  warning('You have written more than two arguments in the
hyphen.')
  return
end
fprintf('WER, case sensitive          WER, case insensitive\n')
fprintf('%6.3f                    %12.8f\n',w(1),w(2))

function d=strd(a,b,k)
% d = strd (a, b, cas) calculates the distance to Levenshtein
```

Noisy	A	B	C	Avg.
Mel	12.42	21.64	16.90	16.98
Gammatone	13.90	22.78	16.07	17.58
FFT	13.26	21.78	15.30	16.78

```
% strings a and b with the aid of the Vagner-Fisher algorithm
  aa=a;
  bb=b;
 if k==2
  aa=upper(a);
  bb=upper(b);
 end
luma=numel(bb); lima=numel(aa);
lu1=luma+1;      li1=lima+1;
dl=zeros([lu1,li1]);
dl(1,:)=0:lima;   dl(:,1)=0:luma;
%Distance
for i=2:lu1
 bbi=bb(i-1);
 for j=2:li1
  kr=1;
  if strcmp(aa(j-1),bbi)
   kr=0;
  end
 dl(i,j)=min([dl(i-1,j-1)+kr,dl(i-1,j)+1,dl(i,j-1)+1]);
 end
end
d=dl(end,end);
```

TABLE I: Word error rates (WERs in %) obtained by the various feature extractors considered in this paper, on the AURORA-2 under clean training conditions.

Clean	A	B	C	Avg.
Mel	8.78	48.61	27.68	28.35
Gammatone	9.44	46.53	24.22	26.73
FFT	9.90	47.03	21.37	26.10

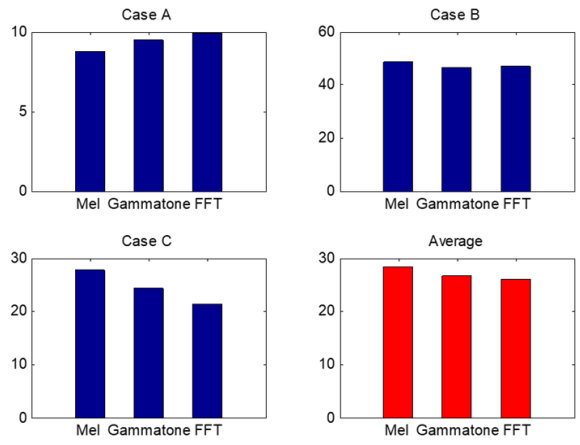


Fig. 7 Word error rates (WER in %) on the AURORA-2 under clean training condition

TABLE II: Word error rates (WERs in %) obtained by the various feature extractors considered in this paper, on the AURORA-2 under multistyle training condition. The lower the WER the better is the performance of the feature

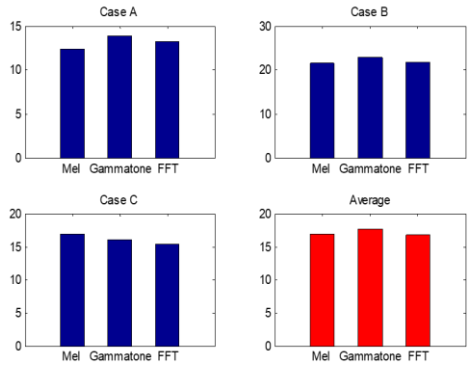


Fig. 8 Word error rates (WER in %) on the AURORA-2 under noisy training condition

Conclusions

Based on research conductedI concluded that direct use of the FFT spectrum offers a more appropriate approach to the perceptual scales rather than the use of filter bank averaging. Direct use of FFT spectral value, or using a Mel or a Gammatone filter bank, takes into account Mel, or Mel, like the frequency scaled frequency cepstral coefficients of Mel (MFCCs), as the features extracted from FFT spectrum values. Computing speech signals using FFT or filter bank spectral features is shown to be more effective in terms of ASR accuracy. I consider the effects of an improved ASR process in the areas discussed to be viable step forward in the field of ASR.

References

[1]Piccone, J. (1992) Signal Modelling Techniques in Speech Recognition. *Texas Instruments*

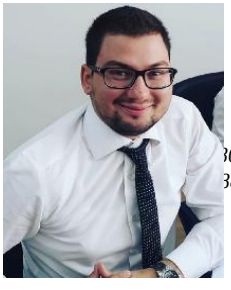
[2]Besacier, L., B, Barnard, E., Karpov, A., Schultz, T. (2014) Automatic speech recognition for under-resourced languages: A survey. Elsevier.

[3]Sainath, T. N, Kingsbury, B., Mohamed, A., Ramabhadran, B. (2013) Learning filter banks within a deep neural network framework. IEEE

[4] Shrawankar, U., Thakare, V. (2013) Techniques for feature extraction in speech recognition system: a comparative study.

[5] Patterson R.D. & Moore, B.C.J. (1986) Frequency Selective in Hearing, chapter Auditory filters and excitation patterns as representations of frequency resolution, pp. 123–177, Academic Press Ltd., London

[6] Parinam, V. N. D. (2013) Spectral analysis methods for automatic speech recognition applications. State University of New York at Binghampton



Saimir Tola was born in Peshkopia, Albania, in 1987. He received the BSc degree in Mathematical Engineering from the Polytechnic University of Tirana, Albania, in 2009 and the MSc (Tech.) and Doctor (Tech.) degrees from Polytechnic University of Tirana, Albania, in 2011, and 2016, respectively. From 2011 until now, he has been working as full-time Lecturer in the field of Numerical Analysis and Signal Processes. His research interest are numerical modeling of Speech Signals, especially Automatic Speech Recognition, interpolation and approximation methods.