

Comparing the Effectiveness/Robustness of Gammatone and LP Methods with the direct use of FFT

Saimir Tola¹, Alfred Daci¹

Polytechnic University of Tirana Faculty of Mathematical and Physical Engineering, Tirane, Albania,¹

Polytechnic University of Tirana Faculty of Mathematical and Physical Engineering, Tirane, Albania,¹

saimir_tola@yahoo.com

Abstract: In this paper we evaluate the growth of Automatic Speech Recognition systems in respect to the various forms of spectral analysis ways used. A straightforward analysis of platter and Gammatone filter banks used for spectral analysis compared with the direct use of FFT spectral values is taken into account. This analysis was supported understanding the effectiveness of existing Automatic Speech Recognition systems that are specifically targeted on platter and Gammatone filter banks compared with FFT spectral values. We discover that warping the FFT spectrum directly, instead of using filter bank averaging, provides an additional precise approximation to the sensory activity scales. Direct use of FFT spectral values are even as effective as using either Gammatone or Linear Prediction filter banks, as long as the feature extracted from the FFT spectral values takes into consideration a Gammatone or platter like frequency scale. Computing speech signals using FFT or filter bank spectral features and utilizing a method supported by a sliding block of spectral features, is shown to be simpler in terms of ASR accuracy.

KEYWORDS: AUTOMATIC SPEECH RECOGNITION, LINEAR PREDICTION, GAMMATONE, FAST FOURIER TRANSFORMATION

1. Introduction

The speech is the basic approach between human beings to communicate and to exchange information with one another. It is only rational to see the importance in researching on this topic. Speech Recognition is a process of altering speech signal to a classification of words by means Algorithm executed as a computer program. The main purpose of speech recognition zone is to developed techniques and systems for speech input to machine based on major advanced in statically modeling of speech, automatic speech recognition today finds widespread application in tasks that require human machine interface such as automatic call processing. [1]. Since the 1960s computer scientists have been researching ways and means to make computers able to record interpret and understand human speech. Throughout the decades this has been a daunting task. Even the most rudimentary problem such as digitalizing (sampling) voice was a huge challenge in the early years. It took until the 1980s before the first systems arrived which could actually decipher speech. Of course these early systems were very limited in scope and power. Communication among the human being is ruled by vocalized linguistic, therefore it is only natural for people to expect speech interfaces with computer seeing the development of the technology now days. So it is necessary to improve computers which can understand spoken words in native language [2].

2. Speech recognition techniques

We can say that the main difficulties about speech recognition are, speaker identification and speaker authentication. Through speaker identifications we can try to guess or identify the person or the voice from a set of known voices.

A speaker identification, is a process in which a sound from an unidentified chatterer is analyzed and compared with speech models of known chatterers. The unknown speaker is identified as the one whose model best matches the input utterance.

Speaker Authentication is the process in which we determine if the chatter claiming to be the actual one is true or not. It is assumed that 'pretenders' are not identified on the system. Also inaccuracy that can occur in speaker identification is the false identification of speaker and the inaccuracies in speaker authentication can be categorized as: (1) false rejections: a true chatterer is rejected as a pretender, and (2) False acceptances: a false chatterer is accepted as a true one [3].

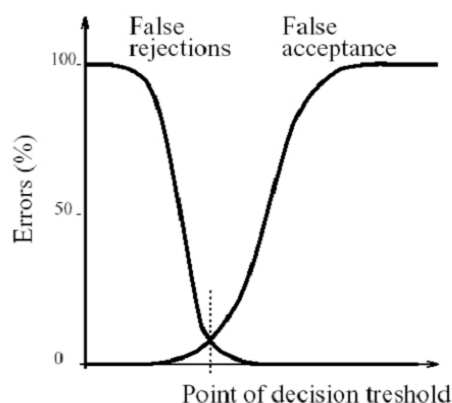


Figure 1. Decision threshold for false rejections and false acceptances. Reproduced from [4]

Table 1: Mechanical properties of selected powder materials Грешка! Источникът на препратката не е намерен. *Note: Values in brackets are valid for heat-treated material.

In most systems for speaker recognition, a distance towards stored speaker's template is computed and is compared with prearranged threshold. If the computed distance is below the threshold the chatterer is verified, otherwise chatterer is rejected as an imposter. The decision threshold is located at the point where the probabilities of both the errors are equal as shown in the Figure 1.

We can categorized speaker recognition methods into text \bar{n} dependent and text \bar{n} independent methods. In case of text \bar{n} dependent methods a chatterer is required to utter a predetermined set of words or sentences. Features of voice are extracted from the same utterance. In case of text \bar{n} independent methods, there is no fixed set of words or sentences and the chatterers do not know that they are being tested. But this methods have their disadvantages. For example if we play a recorded voice of an identified chatter/speaker the system may recognize the chatter as a registered chatter/speaker. Hence, a text \bar{n} stimulated speaker recognition system could be well-thought-out. With the combination of speaker and speech recognition systems and improvement in speech recognition accuracy, the distinction between text \bar{n} dependent and independent applications will eventually be reduced.

2.1. Linear Prediction

The main idea of the filter bank will be combined with the LP model for ASR. Linear prediction-derived amplitudes is outlined as filter bank amplitudes that result from the sampling the LP model of the spectrum rather than the spectrum of the speech signal. Therefore, the LP spectrum is sampled at the relevant filter bank frequencies. The benefit of combining the LP model with filter bank frequencies is that this high resolution model offers spectral estimates that are a lot of stable. Thanks to the spectral smoothing practicality of the LP model a lot of stable parameters to successive stages of the speech process system area unit are created by this mix. However, with the innovations in speech recognition and DSP technologies, the good thing about this mix has adult less over time. To with efficiency sample the spectrum victimization this LP model; the LP model spectral values should be applied directly by computing filter bank amplitudes from the LP model:

$$S_{LP}(f) = \frac{G_{LP}}{\sum_{i=0}^{N_{LP}} a_{LP}(i) e^{-j2\pi(\frac{f}{f_s})i}} \quad (1)$$

Linear prediction (LP) is one in every of the foremost necessary tools in speech analysis. The philosophy behind linear prediction is that a speech sample will be approximated as a linear combination of past samples. Then, by minimizing the ad of the square variations between the particular speech samples and therefore the linearly foretold ones over a finite interval, a singular set of predictor coefficients will be determined [5]. LP analysis decomposes the speech into 2 extremely freeance parts, the vocal tract parameters (LP coefficients) and therefore the glottal excitation (LP residual). It's assumed that speech is created by exciting a linear time-varying filter (the vocal tract) by random noise for unvoiced speech segments, or a train of pulses for voiced speech. Figure A.1 shows a model of vocalization for record analysis [6]. It consists of a time variable filter (z) which is excited by either a quasi-periodic or a random noise supply. The foremost general predictor kind in linear prediction is that the (ARMA) model wherever the speech sample $s(n)$ is modelled as a linear combination of the past outputs and therefore the present and past inputs [7–9]. It will be written mathematically as follows:

$$S(x) = - \sum_{k=1}^p a_k S(n-k) + G \sum_{l=0}^q b_l u(n-l), \quad b_0 = 1$$

2.2 Gammatone Filter Bank

ASR systems include a filter similar to the Mel filter in the context of emulating the human auditory system; the Gammatone filter. The Gammatone filter is a signal processing approximation of the human ear that is designed using a dedicated design of mathematical forms and frequency responses designed to match human physiological experimental results. Like MFCC that was discussed in the preceding section, this feature is usually referred to as GFCC. The Gammatone filter bank was originally introduced by Roy Patterson and colleagues in 1992. Gammatone filters were conceptualized to simply and effectively mimic experimental observations of the mammalian cochlea. The Gammatone filters have a repetitive pole structure leading to an impulse response that is the product of a Gamma envelope. A Gammatone filter is a linear or sequential filter that is depicted by an impulse response. It is a widely used feature in ASR systems. It can be formally represented as follows in the form of an impulse response in the time domain by the following equation: A set of Gammatone filters with different f_c form a Gammatone filter bank. This can be utilized and applied to receive signal characteristics at various frequencies which results in a temporal-frequency presentation which is comparable to the FFT-based short-time spectral analysis. In an effort to mimic human

auditory behavior, the central frequencies of the filter bank are often equally distributed on the Bark scale. [10]

2.3 Fast Fourier Transformation

Within the Fourier transform techniques, the Fast Fourier Transform (FFT) algorithm is the most widespread method. The FFT was invented during the 1960s, and it calculates the frequency representation of a signal of size N in $O(N \log N)$ time. A Fast Fourier Transform (FFT) can be used as a substitute technique of processing the spectrum of the signal. The FFT is a computationally operational deployment of the Discrete Fourier Transform under the limitation that the spectrum is estimated at a discrete set of frequencies that are multiples of $\frac{f_s}{N}$:

These frequencies are called as orthogonal frequencies. The main benefit of the FFT is that it's very fast. It's normal to feature an additional process phase. It's assumed that zones of most vibration or amplitude are given a lot of importance within the human hearing system that are low amplitude zones [11]. Hence, in noisy backgrounds the ground noise inclines to badly influence our estimations of the low amplitude zones of the spectrum during a unbalanced manner. We have a predisposition to lean towards to be dependent on our evaluations of the high amplitude zones of the spectrum. A boundary on the dynamic vary of the spectrum is sometimes required. This low boundary is named the dynamic vary threshold. When we chose a particular threshold from the height within the spectrum, we have an affinity to simply cut or remove the approximations below a specific threshold from the very best purpose inside the spectrum. Threshold algorithmic instruction are used on the spectrum to support the Fourier re-model. For the algorithmic rule, it's required that the spectral values are relatively flat before preparation. Because of the exact circumstance that the spectrum of the audio signal of human speech descends twenty decibel every ten years, having a threshold reinforced on low frequency energy is genuine. Often this may be because of the fact that once rock bottom to highest spectral amplitude diverge division is giant relevant audio energy at higher frequencies can be shattered. FFT puts in order a supportive method of setting a threshold and protective the foremost useful knowledge for process speech signal.

3. Experimental results

To estimate the sustainability and stability of the proposed method in two cases of clean and noisy situations, further to compare with the multiple feature extraction methods, such as LP filter bank, Gammatone filter bank, FFT, the AURORA-2 database is chosen for the experiments. The speech model is trained by different SNR degrees (-5 to 20 dB step -5 dB) of mixed dataset compering clean and noisy data. To produce the noisy data, the clean dataset is extra mixed with three types of noise data, including (1) suburban train, (2) car, and (3) exhibition hall. Totally, 9,880 utterances are yielded for training purposes and split equally into 30 subsets, each SNR subset contains 330 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. 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 covers 9900 utterances and totally 237 600 utterances are generated for testing. For testing the data above us will use the word error rate. We have created these algorithms in Matlab as below, and after the execution we get the data in table 1 and 2.

4. Conclusions

Based on the shown research, we have determined that using the FFT spectrum directly, offers a more exact estimation to the perceptual scales rather than using filter bank averaging. Direct use of FFT spectral values are just as effective as using either LP or Gammatone filter banks. Computing speech signals using FFT or filter bank spectral features is shown to perform better in terms of ASR exactness accept the second case of a noisy background where the LP a little bit better. We consider the effects of an upgraded

ASR process in the areas discussed to be a viable step advancing in the arena of ASR.

Table 1: 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.
LP	10.07	54.61	31.23	31.97
GAMMATONE	11.51	54.33	24.22	40.91
FFT	9.90	47.03	21.37	26.10

Table 2: Word error rates (WERs in %) attained by the numerous feature extractors measured in this paper, on the AURORA-2 under multi style training condition. The lower the WER the better is the performance of the feature

NOISY	A	B	C	Avg.
LP	10.07	54.61	31.23	31.97
GAMMATONE	11.51	54.33	24.22	40.91
FFT	9.90	47.03	21.37	26.10

5. References

- [1] R.Klevansand R.Rodman, "Voice Recognition, Artech House, Boston, London 1997.
- [2] Samudravijaya K. Speech and Speaker recognition tutorial TIFR Mumbai 400005.
- [3] Silveira, M. A., Schroeder, C. P., Paulo, J., Lustosa da Costa, C., De Oliveira, C. D.,
- [4] Trentin, E., Gori, M. (2001) A Survey of Hybrid ANN/HMM Models for Automatic Speech Recognition, Neurocomputing 37(1), pp. 91-126.
- [5] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition. Englewood Cliffs, New Jersey: Prentice-Hall, 1993.
- [6] L. R. Rabiner, Digital Signal Processing. IEEE Press, 1972.
- [7] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., vol. 50, pp. 637-655, Aug. 1971.
- [8] J. Makhoul, "Linear prediction: A tutorial review," Proc. IEEE, vol. 63, pp. 561-580, Apr. 1975.
- [9] B. S. Atal and M. R. Schroeder, "Linear prediction analysis of speech based on a pole-zero representation," J. Acoust. Soc. Am., vol. 64, no. 5, pp. 1310-1318, 1978.
- [10] Alam, M. J., Kenny, P. Dumouchel, P., O'Shaughnessy, D. (2014) Robust speech recognition using warped DFT based cepstral features in clean and multistyle training. IEEE.
- [11] Piccone, J. (1992) Signal Modelling Techniques in Speech Recognition.