Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 3159-0040
Vol. 2 Issue 10, October - 2015

A Comparative Study Of Linear Predictive Analysis Methods With Application To Speaker Identification

Over a scripting programing

Ervenila Musta
Department of Mathematics
Faculty of Mathematics and Physics Engineering
Polytechnic University of Tirana
Tirane , Albania
ervimst@yahoo.com

Vangjush Komini
Department of Mathematics
Faculty of Natural Science
3 RUG Rijksuniversiteit Groningen
Groningen ,Holland
vkomin2@gmail.com

***Abstract*—This paper introduces a generalized formulation of linear prediction (LP), including both conventional and temporally weighted LP analysis methods as special cases. The temporally weighted methods have recently been successfully applied to noise robust spectrum analysis in speech and speaker recognition applications. In comparison to those earlier methods, the new generalized approach allows more versatility in weighting different parts of the data in the LP analysis. Two such weighted methods are evaluated and compared to the conventional spectrum modeling methods FFT and LP, as well as the temporally weighted methods WLP and SWLP. Weighted linear prediction (WLP) is a method to compute all-pole models of speech by applying temporal weighting of the square of the residual signal. By using short-time energy (STE) as a weighting function, this algorithm was originally proposed as an improved linear predictive (LP) method based on emphasising those samples that fit the underlying speech production model well The study compares the performances of SWLP algorithm with the performances of WLP and FFT algorithm .This linear predictive analysis methods are studied and compared from the point of view of robustness to noise and of application to speaker verification with implementation in MATLAB .**

Keywords— Linear Predictive , Weighted Linear Predictive,SWLP, Matlab .

I. INTRODUCTION

Speaker verification technique is becoming a wide area of research. There are several technique and combined method for speech recognition [1] [5]. Accuracy is the biggest concern in every feature recognition method. In order to make the extraction of the features as precise as we can we need to intrude in every single step of this process. Even though the technique is text independent there are several other issue such as noise interference from the background or component parameter mismatch. System is divided in two main benchmark. Fist is feature extraction of the speech signal, then we have feature matching. These are both important, therefore making better system means improving either of them or even both. In our understanding is quite important to make the first step as better as we can. It was quantified important [1] for the technique because of it is impossible to make a decision for the training data unless you provide to the second step comprehensible features. Even with a very accurate matching technique you cannot have reliable decision because you feeded training data might be highly corrupted. Thereby the better your feature extracted are provided to the second step the easier is going to be for making a good decision. Among the most used matching technique two are widely used in different Gaussian Mixture Model GMM [5],[6],[7] and Support Vector Machine SVM [6],[7]. On the other hand, commonly used technique for the second step is Mel_Frequence Cepstral Coefficients MFCCs. This technique has itself some substeps where the features emphasized using Discrete Fourier Transform DFT, or the fast implementation Fast Fourier Transform FFT. This is very important because this is how we get all the information about the intonation of a particular speech signal. For pattern matching we just compare different intonation hence this step should be performed very accurately. Since additive is present in real life implementation there might be needed to do some speech enhancement before we perform feature extraction. This enhancement could be done using some filtering [8] or it could be processed statistically [9]. This require more cost and computation but it is a big strength for the system, and it lowers down interfered data contamination. Present application of MCCF extract features using all-pole mode Linear Prediction LP and it was relatively successful until Weighted Linear Prediction. It was proposed as very competitive feature extracting method. Indeed is gives a better view of speech intonation, however it face some drawback when the signal-to-noise ration fells down under a certain threshold. However this method was optimised even further, providing us more emphasized. In this paper we will present a

comparative research. Instead of WLP we will run Stable Weighted Linear Prediction SWLP. This technique has stable poles hence its performance is way better. Our estimation shows very good power spectral, for the same speech signal. EER is different for different SNR value if we run core implementation. Specifically for SWLP the result is much better compare to the others. For the same SNR we get better EER, consequently the accuracy of this implementation is much higher. In the second section we will introduce different feature extraction process, and result under additive noise. The third is the section goes through the pattern matching technique. Whereas the fourth section describes results and conclusions.

II. FEATURE EXRACTION

A. LINEAR PREDICTIVE MODELS

In speech science, linear predictive methods have a particularly established role, due to their close connection to the source-filter theory of speech production and its underlying theory of the tube model of the vocal tract acoustics. The model provided by LP is especially well-suited for voiced segments of speech, in which AR modelling allows a good digital approximation for the filtering effect of the instantaneous vocal tract configuration on the glottal excitation. The original formulation of WLP, however, did not guarantee stability of all-pole models. Therefore, the current work revisits the concept of WLP by introducing a modified short-time energy function leading always to stable all-pole models. This new method, stabilized weighted linear prediction (SWLP), is shown to yield all-pole models whose general performance can be adjusted by properly choosing the length of the STE window, a parameter denoted by M.Linear predictive speech spectrum modeling [7]assumes that each speech sample can be predicted as a linear combination of *p* previous samples

$$\hat{x} = \sum_{i=1}^{p} a_i x_{n-i}$$

where x_n are the samples of the speech signal in a given short-term frame and $\{a_k\}$ are the predictor coefficients .The number of predictor coefficients *p* is the order of linear prediction .The prediction error is denoted as

$$e_n = x_n - \widehat{x_n} = x_n - \sum_{i=1}^{p} a_i x_{n-i}$$

.Conventional LP analysis minimizes the energy of the prediction error signal

$E_{LP} = \sum_n e_n^2 = \sum_n (x_n - \sum_{i=1}^{p} a_i x_{n-i})^2$

by setting the partial derivatives of E_{LP} with respect to each coefficient a_i to zero .This results in the normal equations [7]

$\sum_{i=1}^{p} a_i \sum_n x_{n-i} x_{n-j} = \sum_n x_n x_{n-j} \quad 1 \le j \le p.$

Although not explicitly written ,the range of summation of *n* is chosen to correspond to the autocorrelation method ,in which the energy is minimized over a theoretically infinite interval ,but x_n considered to be zero outside the actual analysis window [7] .An important benefit of the autocorrelation method is that the LP synthesis model

$$H(Z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

is guaranteed to be stable , i.e ,the roots of the denominator polynomial are guaranteed to lie inside the unit circle [7].

B . WEIGHTED LINEAR PREDICTION (WLP)

Weighted Linear Prediction (WLP) [5] is a generalization of LP analyses .In contrast to conventional LP ,WLP introduces a temporal weighting of the squared residual in model coefficients $\{b_k\}$ are solved by minimizing the energy $E_{WLP} = \sum_n e_n^2 W_n = \sum_n (x_n - \sum_{i=1}^{p} b_i x_{n-i})^2 W_n$ (1)

Where W_n is the weighting function .The Weighting can be used to emphasize the importance of the prediction error in the temporal regions assumed to be less affected by noise ,and de-emphasize the importance of the noisy regions.The WLP model is obtained by solving the normal equations $\sum_{k=1}^{p} b_k \sum_n W_n x_{n-i} x_{n-j} = \sum_n W_n x_n x_{n-j}$ $1 \le j \le p$ (2) .

It is easy to show that conventional LP can be obtained as a special case of WLP :by setting $W_n = d$ for all n ,where $d \neq 0$,d becomes a multiplier of both sides of (2) and cancels out ,leaving the LP normal equations .Typically ,the weighting function W_n in WLP is chosen as the short-time energy (STE) of the immediate signal history [5][6][11][14].
$W_n = \sum_{i=1}^{M} x_{n-i}^2$,

where *M* has previously been chosen close to or equale to the value of *p* [11][14].When compared to conventional spectral modeling method such as FFT and LP,WLP using STE weighting has been recently shown to improve robustness with respect to additive noise in the feature extraction stages of both large vocabulary continous speech recognition [11]and speaker verification [14].

C . STABILIZED METHOD (SWLP)

WLP is not guaranteed to produce a stable all-pol synthesis model
$H(Z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$
(even when using the autocorrelation method ,which in conventional LP always gives a stable model).As a remedy ,a stabilized version of WLP ,called SWLP ,was developed in [6].Although SWLP is stabilized

mainly for synthesis purposes ,it has been found ,like WLP,to be a robust method in the feature extraction stages of speech recognition [6][11] and speaker verification [14] –even surpassing WLP in performance in the latter application .As stated in section 2.1.1 ,the WLP normal equations can be rewritten as :

$\sum_{i=1}^{p} b_i \sum_n Z_{n,i} x_{n-i} Z_{n,j} x_{n-j} = \sum_n Z_{n,0} x_n Z_{n,j} x_{n-j} \quad 1 \le j \le p$ (3)

Where $Z_{n,j} = \sqrt{W_n} \ \ for \ 0 \le j \le p$. As shown in [6] (using a matrix-based formulation),model stability is guaranteed if the weights $Z_{n,j,}$ are instead ,defined recursively as $Z_{n,0} = \sqrt{W_n}$ and $Z_{n,j} = \max(1, \frac{\sqrt{W_n}}{\sqrt{W_{n-1}}}) Z_{n-1,j-1} \quad 1 \le j \le p.$

Substitution of these values in equation (3) gives the SWLP normal equations .

III .IMPLEMENTATION

This is the step where our research is mostly focused on. As it is stated on the introduction section this step is where all speech features are extracted from the speech signal of time domain. Even further, this is presented with MFCCs method using different windowing function for the periodogram.. In computer simulation we will use FFT [8] instead as lower time complexity. x[n] is assumed to be zero outside the interval [0,N-1]. On the other hand, linear prediction from the above section is based on the idea that the upcoming data point can be predicted from the previously data point. It is characterized from the order of prediction. Later on, instead of using a simple linear prediction a better version with a weighted function was proposed. Unlike simple linear prediction here we will try to minimize the product of error function with this weighting function. This is in time domain weighted function.

Our research is mainly focused on Stabilised Weighted Linear Prediction. This reveal way better feature spectrum and the complexity is not increased drastically. Since the stability for the WLP is not guarantee it is needed to do some additional application do make sure we have all the pole model within the unite circuit. If we run the above method over a discrete data speech signal in time domain, they yield different result. Differences between three plots reveal a significant improvement from the simple Linear Prediction, up to Stable Weighted Linear Prediction. Value of the function yield the amplitude of the feature speech signal. The information we get from LP case is not that detailed compare to the other method. Plot from LP is smooth thereby it doesn't reveal that much information. This is so because it doesn't give a big picture of the feature differences between two different frequency value. WLP gives us a better information about the features. From the plots we can see that features are more detailed because the plot is more hilly, and hence you know better how features differ in different frequencies.

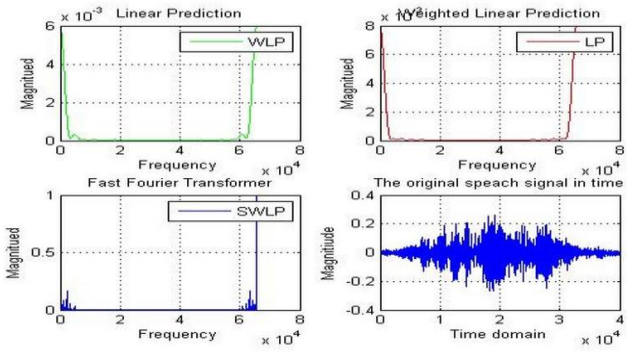

Fig 1. *Comparison of SWLP versus WLP,LP*

Moreover, since in real life application we have the presence of the noise therefore we need to see the effect of noise. Below is a simulation of different speech signal under different additive noise.

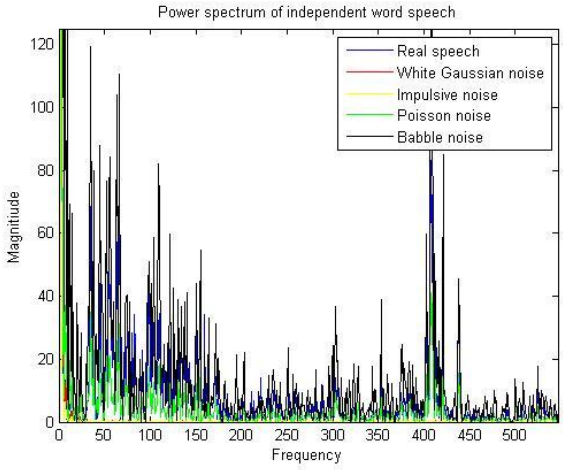

Fig 2. *Power spectrum of independent word speech*

A . PATTERN MATCHING.

After the feature extraction and MFCC, next step is pattern matching. This is also an important step for the verification result and it could be done through several method [6].

- Nearest Neighbour k-NN
- Bayes` Classifier
- Artificial Neural Networks ANN
- Gaussian Mixture Model GMM
- Support Vector Machine SVM

All the above alternative could be one of the implementation for pattern matching. Due to the large number of paper published for GMM during the last past years GMM has become very dominant approach for text-independent verification. GMM with universal background model is implemented widely for speech verification. Universal Background Model is a model for biometric verification system , for person-independent features to compare against a model of person-specific feature characteristics when making an accept or reject decision. In speaker verification, the UBM is a speaker-independent GMM trained with speech samples from a large set of speakers to represent general speech characteristics. This approach goes through the following benchmark:

B . LIKELIHOOD RATIO DETECTION

Overall in this method, the goal for a given speech segment Y and a hypothesis speech segment S is to determine whether Y and S are coming from the same source. The only restriction is that we will assume that Y contains speech from only one speaker, this is also known as single speaker detection. This can be stated as a simple hypothesis testing between
H0: Y and S are from the same source
H1: Y and S are from different source
The likelihood ratio is to decide the optimum test between these two hypothesis given by:

$$\frac{P(Y|H0)}{P(Y|H1)} = \begin{cases} \geq \theta \; accept \; H0 \\ < \theta reject \; H0 \end{cases}$$

where p(Y|Hi), i=1,2 is the probability density function for the hypothesis Hi evaluated for the observed speech segment Y, referred to as the likelihood for the hypothesis Hi, give the speech segment. The decision threshold H0 is . Defining the value of p(Y|H1) and p(Y|H2) is challenging as well. One way for doing this is described in the figure below

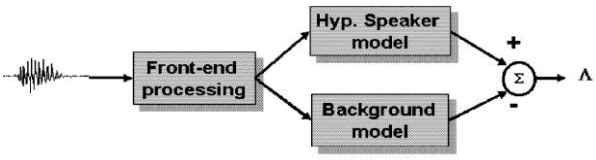

Fig 3 Likelihood ratio-based speaker detection system

As front-end processing we could employ linear filtering of the hypothesis speech segment data vector X={x[1], x[2],....x[T]} at discrete time domain T={1,2,...T}. X is the feature vector.

C . GAUSSIAN MIXTURE MODEL

Selection of the actual likelihood function p(X|), is important, since it depends on the features being used and the specific application. For the text-independ, where there is no prior knowledge about what speaker is going to say, GMM is the most successful likelihood function.
For a D-dimensional feature vector, x, the mixture density used fro the likelihood function is defined as

$$p(\boldsymbol{x} \mid \lambda) = \sum_{i=1}^{M} w_i \, p_i(\boldsymbol{x})$$

where:

$$p_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\mu_i)'(\Sigma_i)^{-1}(\boldsymbol{x}-\mu_i)\right\}$$

The restriction for this is that i=1Wi=1where i=1…..M. The parameters of the density model are ={i,i,i}The GMM can be viewed as a hybrid between parametric and nonparametric estimation. The advantages of using a GMM as likelihood function are that it is computationally inexpensive, is based on well-understood statistical model and, for text-independent, is insensitive to the temporal aspects of the speech, modeling only the underlying distribution of acoustic observation from a speaker.

D . Front-End Processing

The speech is segmented into frames by 20-ms window progressing at a 10-ms frame rate. The speech detector discards 20-25 % of the signal. Next mel-scale cepstral feature vectors are extracted from the speech frames. The mel-scale cepstrum is the discrete cosine transform of the log spectral energies of the speech segment Y. The spectral energies are calculated over logarithmically spaced filter with increasing bandwidth. Delta cepstra are computed using a first order orthogonal polynomial temporal fit over 2 feature vectors.

IV . RESULT

We need to describe different classification error and explain how the quality of two system can be compared objectively. A pattern that is going to be verified is matched against the known template, yielding either a score or a distance describing the similarity between the pattern and the template. In order to have a reliable result, the similarity gas to exceed a certain level. Unless the level is reached, the pattern is rejected. However the classification threshold is chosen, some classification errors occur. You can choose the threshold such high, that no impostor scores will exceed this limit, consequently no patterns are falsely accepted. Unlikely all the patterns with score lower that the highest impostor score are falsely rejected. You can choose the threshold such low that no client patterns are falsely rejected graning this some impostor patterns are falsely accepted. Thereby if you choose the threshold somewhere between those two points, both false rejection and false acceptance occur. The threshold depending fraction of the falsely accepted patterns divided by the number of all impostor patterns is called **False Acceptance Rate (FAR).** FAR is one if all impostor patterns are falsely accepted and zero, if none of the impostor patterns is falsely accepted. The fraction of the number of rejected client patterns divided by the total number of the client patterns is called **False Rejection Rate (FRR).** It is one if all impostor patterns are falsely rejected and zero, if none of the impostor patterns is falsely rejected. At the point where **FAR** and **FRR** become equal, is called **Equal Error Rate (EER).** This can be used to give a threshold independent performance measure. The lower the **ERR** the better is the, system's performance, as the total error rate is the sum of **FAR** and **FRR** at the point of **ERR**.
Our research is based on measuring different EER for different feature extraction technique, under different signal-to-noise rate.

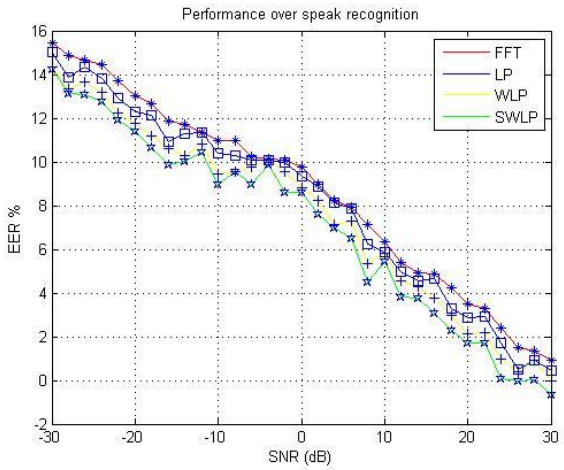

Fig 3.Performance over speak recognition

V .CONCLUSION

In this paper ,we discussed the performance of the Stabilized Weighted Linear Prediction .This method applies temporal weighing on the square of the residual signal and thus emphasizing the samples of the high energy ,which typically belong to closed phase interval during phonation .

VI .REFERENCES

[1] P. Strobach, Linear Prediction Theory-A Mathematical Basis for Adaptive Systems, Springer-Verlag, 1990.
[2] S. Haykin, Communication Systems, 4th Ed., John Wiley & Sons,2001
[3] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, “Speaker verification using adapted Gaussian mixture models,” Dig. Sig. Proc., vol. 10, no. 1, pp. 19–41, Jan. 2000.
[4] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P.Alku, “Temporally weighted linear prediction features for tackling additive noise in speaker verification,” IEEE Sig. Proc. Lett., vol. 17, no. 6, pp. 599–602, June
[5] Handbook of speech recognition Benesty, Jacob; Sondhi, M. M.; Huang, Yiteng (Eds.) 2008
[6]Pattern Recognition and Machine Learning C. Bishop,1 Feb 2007, ISBN-10: 0387310738, 2-nd edition
[7]Makhoul,J ,Linear prediction a tutorial review, Procceding of the IEEE.
[8] Pattern Classification Richard O. Duda, Peter E. Hart, David G, November 9, 2000 | ISBN-10: 0471056693,2-nd edition
[9] Understanding Digital Signal Processing (2nd Edition) by Richard G. Lyons
[10]Fundamentals of Statistical Signal Processing, Volume III: Practical Algorithm, Development April 5, 2013 | ISBN-10: 013280803X .
[11] Pohjalainen ,J,Kallasjoki ,H,Plomaki ,K,J,Kurimo ,M and Alku, P., Weighted Linear Prediction for speech Analysis in Noisy Condition, in Proc .Interspeech ,Brighton ,UK ,2009.
[12]Saedi ,R,Pohjalainen ,J,Kinnunen ,Tand Alku ,P , Temporally Weighted Linear Prediction Features for Tackling Additive Noise in Speaker Verification , IEEE Signal Processing Letters 17(6)2010.