Perceptual Time Varying Linear Prediction Model for Speech Applications

Oron Gamliel and Ilan D. Shallom
Department of Electrical and Computer Engineering
Ben Gurion University of the Negev, Beer Sheva, 84105, ISRAEL
gamlielo@gmail.com, ilan.shallom@audiocodes.com

ABSTRACT

A new perceptual time varying model for non–stationary analysis of speech signals is presented. Some researches have already shown that the Time Varying Linear Prediction Coding (TVLPC) model that was applied to speech signals increases the recognition performance of Automatic Speech Recognition (ASR) systems. This improvement has been achieved due to the incorporation of the speech dynamics information in the model.

Another work, Perceptual Linear Prediction (PLP) analysis of speech**,** has shown that a modified estimation of the Auto Correlation Function (ACF) of stationary speech frame yields major improvement to the recognition rate.

The presented model, Perceptual Time Varying Linear Prediction (PTVLP) analysis of speech, adopts the perceptual concepts, of how to estimate the ACF, into the TVLPC model. This research shows that the proposed PTVLP model is more accurate, robust to noise and achieves better recognition rates than PLP and TVLPC over wide SNR range.

Index Terms— Auto Regressive, HMM, TVLPC, PLP, PSD

1. INTRODUCTION

The PTVLP model is actually a combination of two well known speech analysis models, the Time Varying Linear Predictive Coding (TVLPC) ([1] [2]) and the Perceptual Linear Predictive (PLP) [3]. In practice, the PTVLP model adopts the advantages of the PLP into the TVLPC.

Auto Regressive (AR) filtering is widely used in order to predict the next sample based on previous samples. During the optimization process, for finding the optimal AR filter coefficients (in the MMSE sense), an estimation of the Auto Correlation Function (ACF), or equivalently, the Power Spectrum Density (PSD) of the speech frame, is needed.

This estimation was found to be inconsistent with the perception of human hearing (which is considered as the upper bound classifier). First, the ACF (or PSD) is estimated with equal bin widths around each frequency over the analysis band, where the natural human hearing is less sensitive to minor frequency derivations as the frequency increases. Hence, in order to model better the perception of human hearing, the analysis band needs to be divided to perceptual sub bands. Second, it was found that the perception of human hearing has a different sensitivity to the amplitude levels in each such sub band. Therefore, there is a need to enhance the amplitudes of the estimated PSD per each sub band.

In addition, our AR filter has coefficients that we allow them to slightly vary with time in order to earn representation of the speech dynamics. Both actions, the perception and the time varying, produce better representation of speech and therefore increase the recognition rate. The paper is organized as follows: Section **2** describes the implementation details of the PTVLP model. In section **3**, experiments results of the model are depicted. Section **4** contains the conclusions.

2. THE PTVLP MODEL

Figure 1 depicts the main stages of the PTVLP:

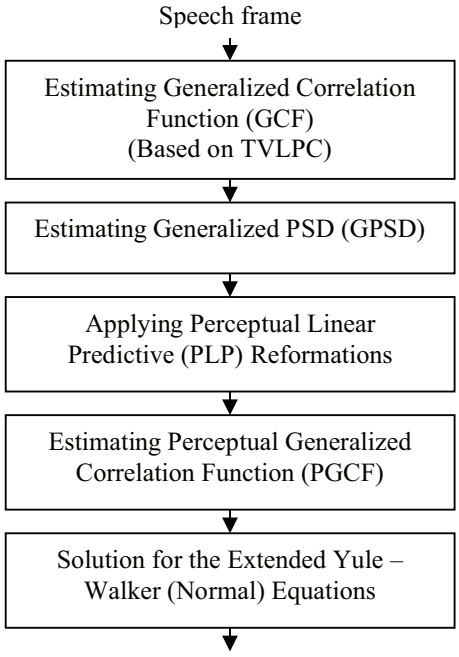

Figure 1: The PTVLP model

As can be seen in Figure 1, we first estimate the General Correlation Function (GCF) of the speech frame. Then, the

Generalized PSD (GPSD) of the speech frame is obtained by applying Fourier transform to the GCF.
The GPSD is perceptually reformatted to yield the Perceptual GPSD (PGPSD). Finally, we substitute the Perceptual GCF (PGCF), which is achieved by applying inverse Fourier transform to the PGPSD, in the Extended Yule – Walker (Normal) equations yielding the PTVLP coefficients.

2.1. Time Varying Linear Prediction Coding

As has been stated in [1], the speech frame signal $s[n]$ at time n, is expressed as a linear combination of the past P samples and the inaccessible input $u[n]$, i.e.,

$$s[n] = -\sum_{i=1}^{p} a_i[n]s[n-i] + Gu[n] \quad (1)$$

A common used constraint on the time varying coefficients, $a_i[n]$, is to model them by a linear combination of some known basis functions, i.e.,

$$a_i[n] = \sum_{k=1}^{Q} a_{ik} f_k[n-i] \quad (2)$$

Where G denotes gain input, $f_k(n)$ denotes the k[th] basis function, $Q+1$ denotes number of basis functions and a_{ik} are the basis functions weights.
Substituting Equation (2) into (1) will lead to the predictor equation:

$$\hat{s}[n] = -\sum_{i=1}^{P} \left(\sum_{k=0}^{Q} a_{ik} f_k[n-i] \right) s[n-i] \quad (3)$$

The prediction error is:

$$e[n] = s[n] - \hat{s}[n] \quad (4)$$

The usual used criterion of optimality for the coefficients is the minimization of the total squared error, defined as:
(5)

$$E_{MSE} = \sum_n e^2[n] = \sum_n \left(s[n] + \sum_{i=1}^{P} \left(\sum_{k=0}^{Q} a_{ik} f_k[n-i] \right) s[n-i] \right)^2$$

Minimization of this error with respect to each coefficient, a_{ik}, will lead to set of equations that are called the Extended Yule – Walker (Normal) equations:
(6)

$$\sum_{i=1}^{P} \sum_{k=0}^{Q} a_{ik} r_{kl}[i-j] = -r_{0l}[0-j] \quad 1 \le i, j \le P, \;\; 0 \le k, l \le Q$$

Where $r_{kl}[m]$ denotes the Generalized Correlation Function (GCF) and defined as:

$$r_{kl}[m] = \sum_n f_k[n]s[n]f_l[n+m]s[n+m] \quad (7)$$

We will denote the set of received optimal coefficients, $\{a_{ik}\}$, as the TVLPC optimal coefficients of the speech frame $s[n]$.

2.2. Estimate Generalized PSD (GPSD)

In order to use the benefits of the perceptual reformations on the speech spectrum, we need to obtain the Power Spectrum Density (PSD) of the speech frame, $s[n]$.
It can be seen that the yielded GCF, $r_{kl}[m]$ (Equation (7)), is a correlation function of the speech frame $s[n]$, multiplied by the deterministic basis functions $f_k[n]$ and $f_l[n]$.
According to the Weiner – Kchinchin theorem, the PSD of signal is the Fourier transform of its autocorrelation function. In general, we can define the Generalized PSD (GPSD) of a signal as the Fourier transform of its GCF.
Thus, we will denote the GPSD as the Fourier transform of the GCF of the signal $s(n)$:

$$\therefore GPSD = P_{kl}(\omega) = F\{r_{kl}(t)\} \quad (8)$$

Where $F\{\cdot\}$ denotes the Fourier transform. In the discrete case, the Discrete time Fourier Transform (DFT) will replace the Fourier transform. Hence:

$$P_{kl}[q] = DFT\left\{r_{kl}[m]\right\} = \sum_{m=0}^{N-1} r_{kl}[m] e^{-j2\pi qm/N} \quad (9)$$

Where $DFT\{\cdot\}$ denotes the Discrete time Fourier transform. It can be shown that substitution of Equation (7) into Equation (9) yields:

$$P_{kl}[q] = DFT\left\{s[n]f_k[n]\right\} \cdot DFT^*\left\{s[n]f_l[n]\right\} \quad (10)$$

Where * denotes the complex conjugate.
One can see that if both basis functions are equal, i.e., $f_k[n] = f_l[n] = f[n]$, we actually get the PSD estimation of $s[n]f[n]$.

2.3. Perceptual Linear Predictive Reformations

Perceptual reformations are applied to the estimated GPSD in order to model the perception of the human hearing.
These perceptual reformations include three main steps that are described in the subsections 2.3.1, 2.3.2 and 2.3.3 below.

2.3.1. Critical Band Analysis

The GPSD has to be warped differently along its frequency axis because of the fact that the spectral resolution of human hearing decreases with frequency beyond about 800Hz. In addition, the spectral resolution of the GPSD should be reduced in order to simulate the natural hearing system. In order to do so, we have used a set of 17 bark – scale auditory shape filters and Equation (11) depicts the formal warping formula:

$$\Theta_{kl}(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P_{kl}(\Omega - \Omega_i)\Psi(\Omega) \quad (11)$$

Where:

- $P_{kl}(f)$ denotes the GPSD of the speech frame when using basis functions $f_k[n]$ and $f_l[n]$.
- $\Omega(f) = 6\ln(\frac{f}{600} + ((\frac{f}{600})^2 + 1)^{0.5})$
- $\Psi(\Omega) = \begin{cases} 0 & \Omega < -1.3, \Omega > 2.5 \\ 10^{2.5(\Omega+0.5)} & -1.3 \le \Omega \le -0.5 \\ 1 & -0.5 < \Omega < 0.5 \\ 10^{-1.0(\Omega-0.5)} & 0.5 \le \Omega \le 2.5 \end{cases}$
- $\Theta_{kl}(\Omega_i)$ is the auditory shaped GPSD.

2.3.2. Equal Loudness Preemphasis

Since the non – equal sensitivity of the human hearing to amplitudes at different frequencies, there is a need to simulate an equal – loudness curve.

Psycho acoustic experiments have found that such step will emphasize the amplitude levels resolution at the frequency ranges of 0 - 400Hz and 1200 – 3100Hz.

The formula that was used in order to apply it to the auditory shaped GPSD, $\Theta_{kl}(\Omega(f))$ or formally:

(12) $\Xi_{kl}(f) = E(f) \cdot \Theta_{kl}(\Omega(f))$, Where:

- $E(f) = \left(\frac{f^2}{f^2 + 1.6 \cdot 10^5}\right)^2 \bullet \left(\frac{f^2 + 1.44 \cdot 10^6}{f^2 + 9.61 \cdot 10^6}\right)$
- $\Theta_{kl}(\Omega(f))$ denotes the auditory shaped GPSD.
- $\Xi_{kl}(f)$ denotes the loudness preemphasized auditory shaped GPSD.

2.3.3. Intensity Loudness Conversion

The final perceptual step is the cubic root amplitude compression. It takes the loudness preemphasized auditory shaped GPSD and yields the third root of it. This operation is an approximation to the Steven's power – law of hearing [3]. It simulates the non – linear relation between the intensity of sound and its perceived loudness, formally:

(13) $T_{kl}(f) = \Xi_{kl}(f)^{0.33}$

This step finalizes the perceptual reformations of the GPSD. $T_{kl}(f)$ actually denotes the Perceptual GPSD (PGPSD) of the speech frame and will be used to obtain the Perceptual GCF (PGCF) of the speech frame.

Figure 2 depicts an example of how the perceptual reformations affect the GPSD. A 50 msec speech frame, sampled at 8 KHz, from the word zero was used. The GPSD was created by using 1024 points FFT. All GPSDs are displayed on dB scale, versus 0 – 4 KHz.

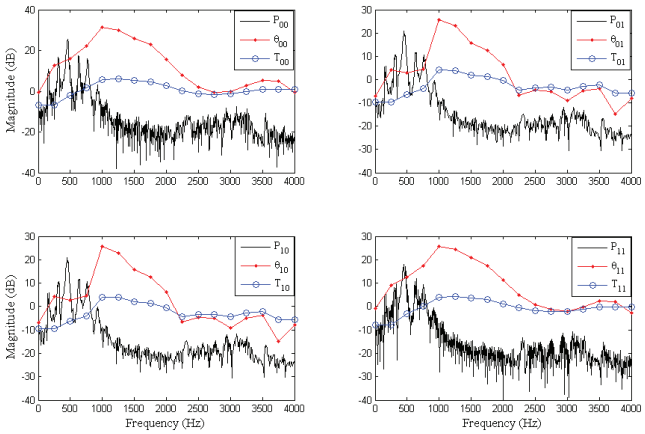

Figure 2: The Perceptual Effect on GPSD from the word "Zero" using 2 basis functions "Polynomials"

$P_{kl}(f)$ in each of the subplots denotes the GPSD (flat line). The high spectral resolution can be easily seen.

$\Theta_{kl}(f)$ denotes the auditory shaped GPSD (line with dots). The decreasing of the spectral resolution in comparison to the GPSD is well seen.

Finally, $T_{kl}(f)$, denotes the Perceptual GPSD (PGPSD) (line with circles). It holds the Equal Loudness Preemphasis and the Intensity Loudness Conversion. It is clearly seen that $T_{kl}(f)$ is quiet flatten over the analysis band (simulates the Equal Loudness Conversion) and its dynamic range is significantly smaller than the GPSD (simulates the Intensity Loudness Conversion).

2.4. Perceptual Generalized Correlation Function

Prior to the obtaining of the PTVLP features, we need to obtain the Perceptual GCF (PGCF). It is substituted in the Extended Yule-Walker normal equations instead of the GCF. PGCF can be easily obtained by applying the inverse Fourier transform to the last obtained PGPSD. i.e.,

(14) $PGCF \stackrel{\therefore}{=} C_{kl}(t) = F^{-1}\{T_{kl}(\omega)\}$

Where $C_{kl}(t)$ denotes the PGCF and $T_{kl}(f)$ denotes the PGPSD.

In the discrete case, the Inverse DFT (IDFT) replaces the inverse Fourier transform. The discrete form of Equation (14) is:

(15) $C_{kl}[m] = IDFT\left(T_{kl}[q]\right) = \frac{1}{N}\sum_{q=0}^{N-1} T_{kl}[q] e^{j2\pi qm/N}$

2.5. Obtaining PTVLP coefficients

This step finalizes the PTVLP analysis of speech. The appropriate lags of the PGCFs were used to solve the Extended Yule Walker (Normal) equations (6) in order to get the optimal (in perceptual MMSE sense) coefficients of the PTVLP model. The extended LWR algorithm was found useful to solve the Yule – Walker (Normal) equations [5].

These coefficients are further used as the observation vector of the correspondent speech frame for recognition purpose.

3. EXPERIMENTS RESULTS

The experiments were performed for isolated words recognition. The speech database that was used for these evaluations is a part of the Texas Instruments DIGIT (TIDIGIT). The database is consisted of the ten English language digits (0 – 9). Each digit has 448 repetitions, half male, half female, with sampling rate of 8 kHz. Each digit utterances were divided to 224 for Training and 224 for Testing. Hidden Markov Model (HMM) [6] was used to model each digit and the observations per each speech frame were the PTVLP coefficients of it. The well known HMM Tool Kit (HTK) [7] was used to train and test the HMMs per digit. Left to right HMMs were used as in general in speech modeling.

Synthetic WGN was added to the clean speech utterances in order to examine the recognition performance in the presence of noise.

Baseline parameters for three different features extraction models: PTVLP, PLP and TVLPC are depicted in Table 1. Equal features rate was selected in order to fairly compare the three techniques.

PARAMETER	PTVLP/TVLPC	PLP
Model Order	5	5
Frame Length	50 msec	25 msec
Frames Overlap	60 %	60 %
Frame Rate	20 msec	10 msec
Features Length	10	5
Features Rate	500 Coeffs/Sec	500 Coeffs/Sec
Basis Functions	Power	Power
Number Of Basis Functions	2	2
Tested SNRs	0 - 30dB	0 - 30dB

Table 1: Baseline Parameters for PTVLP, TVLPC and PLP techniques

Figure 3 depicts the recognition rates versus SNR of the three features extraction models (PTVLP, PLP and TVLPC). For the PTVLP and PLP, we have also implemented additional experiment, by applying a simple, Wiener based, noise reduction method. It can easily seen that the PTVLP model, when using this simple noise reduction method, achieves better recognition rates than all other models from about 2dB SNR and above.

This result assures our basic assumption that the addition of the perceptual principles to the time varying analysis of the speech will increase the model accuracy and therefore, will achieve better recognition results than the two models it is based on, the TVLPC and PLP.

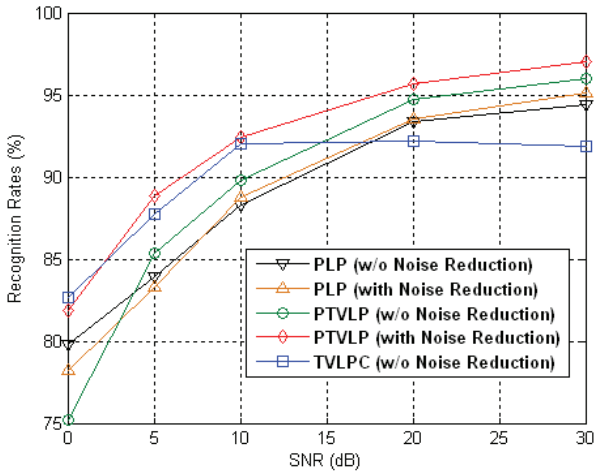

Figure 3: Recognition Rates of PTVLP, PLP and TVLPC in WGN Environment, with and without Noise Reduction

4. CONCLUSIONS

A new perceptual time varying model for non–stationary analysis of speech signals has been described. This model combines the advantages from the TVLPC model and the PLP model in order to increase the agreement of the speech model to the natural human hearing. The PTVLP has achieved better recognition rates than these of the TVLPC and PLP over wide SNR range. The reason for that is the incorporation of the perceptual and dynamic information in the model.

5. REFERENCES

[1] M. G. Hall, A. V. Oppenheim and A. S. Willsky, "*Time – Varying Parametric Modeling of Speech*", Signal Processing 5, pp. 267 – 285, 1983.

[2] I. D. Shallom, "*Non – Stationary Analysis of Speech Signals*", Thesis for Doctor of Philosophy, Ben – Gurion University, Israel, March 1990.

[3] H. Hermansky, "*Perceptual Linear Predictive (PLP) Analysis of Speech*", Journal of the Acoustics, Soc. Am, pp. 1738 – 1752, April 1990.

[4] J. Makhoul, "*Linear Prediction: A Tutorial Review*", Proc. IEEE, Vol. 63, No. 4, April 1975.

[5] R.A. Wiggens & E. A. Robinson, "*Recursive Solution to the Multichannel Filtering Problem*", Journal of Geophysical Research, Vol.70, No. 8, April 1965.

[6] L. R. Rabiner, "*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*", Proceedings of the IEEE, Vol. 77, No.2, February 1989, p. 257-286.

[7] S. J. Young, J. Jansen, J.J Odell, D.G Ollason and P.C. Woodland, "*The HTK book Manual",*3.1, Cambridge University Engineering Department and Entrophic, Cambridge Research Laboratory, Cambridge, UK, December, 2001.