PERCEPTUAL TIME-VARYING MODELLING OF SPEECH SIGNALS FOR ASR AND COMPRESSION APPLICATION

Amir Leibman and Ilan D. Shallom

Electrical Engineering Department, Ben-Gurion University
P.O.Box 653, 84105 Beer-Sheva, Israel
Phone: + (972) 544591577, Email: leibman@bgumail.bgu.ac.il
Phone: +(972) 545750861, Email: shallomi@ee.bgu.ac.il

ABSTRACT

Perceptual audio coders and Automatic Speech Recognition (ASR) systems are commonly based on short-time analysis. This paper presents a generalized model for time-varying coefficients based on psychoacoustic properties of the human ear. The proposed model is evaluated in the framework of speaker independent speech recognition using Hidden Markov Models (HMM). The generalized model is compared to the traditional most popular MFCC. The comparison is made with respect to the models baud rate and the total error rate measured in an extensive Speech recognition experiment. The recognition based on the well established speech recognition development environment, the HTK and using the TIDIGIT as the evaluation database. The time varying model achieves better recognition rate in comparison to MFCC, while the proposed model baud rate is about one third of the baud rate that is used in the case of MFCC. In addition, a preliminary evaluation of the model robustness to noise was carried out and is presented.

1. INTRODUCTION

The human ear performs better than today's best ASR/keyword spotting systems. Thus, the main assumption is that imitation of human ear inner process would improve this kind of system. The goal of this research is to find a spectral time-varying representation based on the psychoacoustic properties of the human ear which models the dynamic along the static part of the speech signal.

Perceptual Linear Prediction (PLP) [1] has already shown better performance in terms of recognition results and baud rate than the traditional LPC algorithm, hence the model is based on the psychoacoustic model.

The paper is organized as follows. The psychoacoustic model is described in section 2. Section 3 contains the time-frequency psychoacoustic model description. Section 4 discusses the final results and Section 5 deals with the conclusion.

2. THE PSYCHOACOUSTIC MODEL

In this paper, the definition used for time-varying spectrum is a spectrogram. Let, $x[j],\ \ 0 \le j \le J-1$, be discrete time non-stationary signal of interest. x[n] is divided into M overlapped frames multiplied by Hamming window, justifying the quasi-stationary assumption.

$$\{x_m[n],_{n=1..N}\}_{m=1..M}.$$

Let the spectrum of $x_m[n]$ be $S_m[k],_{k=1..K}$.

Where, K represents the total amount of bands in the discrete frequency domain. The spectrogram of x[n] is defined as

$$S_x[m,k] = S_m[k],\ \ _{1\le m\le M,0\le k\le K-1}.$$

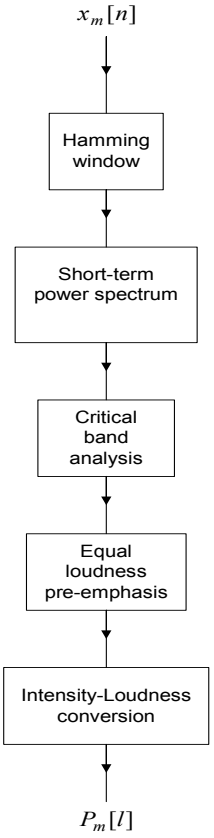

Figure 1: flow chart of the psychoacoustic process

Figure 1 describes the flow chart of the psychoacoustic model. It's input is the short-time frame $x_m[n]$ and the output is $P_m[l]$. Where, $P_m[l],_{l=1..L}$ is defined as bark-scaled psychoacoustic power spectrum density. L defined as the number of critical bands.

2.1 Critical band analysis

The supplied psychoacoustic model is based on Hermanskey's [1] work but analyzed and presented in a different aspect. It differs from PLP's by the critical band analysis. The transform to bark scale referred as a filter bank with resemblance to the Mel filter bank.

The power spectrum is warped onto a bark scale using the following approximation [1]:

$$\Omega(\omega_k) = 6\ln\left(\frac{\omega_k}{1200\pi} + \sqrt{1+\left(\frac{\omega_k}{1200\pi}\right)^2}\right), \qquad (1)$$

where

$$\omega_k = 2\pi\frac{k}{K}f_s,\ \ \omega_k\ \ in\ [\frac{rad}{\sec}]. \qquad (2)$$

In order to reduce the spectral resolution the spectrum of the transformed signal is filtered by a filter bank. Each filter represents a critical band as a bandpass.

Let $\phi[k]$ be defined as

$$\phi[k] = \begin{cases} 0 & ,\Omega(k) < -2.5 \\ 10^{-((-\Omega(k))-0.5)} & ,-2.5 \le \Omega(k) \le -0.5 \\ 1 & ,|\Omega(k)| \le 0.5 \\ 10^{2.5(-\Omega(k)+0.5)} & ,0.5 \le \Omega(k) \le 1.3 \\ 0 & ,1.3 < \Omega(k) \end{cases}. \qquad (3)$$

In (4) the l'th critical band filter is given

$$\phi_l[k] = \phi[k - k_l]. \qquad (4)$$

Each critical band is centred on frequency index k_l matching the frequency in fixed bark value *l*. The bark-scaled power spectrum density is given

$$P_{Bm}[l] = \frac{\sum_{k=0}^{K-1} S_m[k]\phi_l[k]}{\sum_{k=0}^{K-1}\phi_l[k]}, \quad 1 \le l \le L. \qquad (5)$$

Where, $P_{Bm}[l]$ is the power spectrum density measured in the l's, bark scaled, critical band at the instant m.

2.2 Loudness emphasis

Equal loudness emphasis is needed to compensate for the non-equal perception of loudness at different frequencies. The equal-loudness curve is given in (6).

$$E(\omega) = \frac{(\omega^2 + 56.8x10^6)\omega^4}{(\omega^2 + 6.3x10^6)^2(\omega^2 + 0.38x10^9)}. \qquad (6)$$

2.3 Intensity-loudness law

An approximation to the power law of hearing simulates the nonlinear relationship between the sound intensity and the perceived loudness.

$$P_m[l] = (P_{B_m}[l]E(l))^{0.33}. \qquad (7)$$

3. THE TIME-FREQUENCY SPECTRA MODEL

$\mathbf{P}$ is the bark-scaled spectra matrix at the output of the intensity loudness conversion.

$$\mathbf{p} = \begin{pmatrix} p_1[1] & p_1[2] & \cdots & p_1[L] \\ p_2[1] & p_2[2] & \cdots & p_2[L] \\ \vdots & \vdots & \ddots & \vdots \\ p_M[1] & p_M[2] & \cdots & p_M[L] \end{pmatrix}_{MxL}. \qquad (8)$$

The final spectra matrix $\mathbf{P}_L$ defined in log-scale,

$$\mathbf{P}_L = \log_{10}(\mathbf{P}). \qquad (9)$$

Where each component $P_m[l]$ in the matrix $\mathbf{P}$ becomes $P_L[m,l] = \log_{10}(P_m[l])$ in the matrix $\mathbf{P}_L$. The proposed approach models the time-frequency spectra matrix as a linear span of basis functions

$$\hat{P}_L[m,l] = \sum_{i=1}^{T}\alpha_i f_i[m,l]. \qquad (10)$$

Where $\{f_i[m,l]\}_{i=1}^{T}$ is a set representing the basis functions. Complexity considerations led to linear combination of basis functions. Non-linear approximation would be more complicated in a matter of computation complexity.

The coefficients vector is the variable upon which minimization of the error criteria function $J(\boldsymbol{\alpha})$, take place (Least Square Error):

$$J(\boldsymbol{\alpha}) = \sum_{m=1}^{M}\sum_{l=1}^{L}(P_L[m,l] - \sum_{i=1}^{T}\alpha_i f_i[m,l])^2. \qquad (11)$$

The rearrangement of matrix $\mathbf{P}_L$ in a vector representation is defined in (12).

$$\mathbf{p} = \begin{pmatrix} P_L[1,1] \\ \vdots \\ P_L[1,L] \\ P_L[2,1] \\ \vdots \\ P_L[M,L] \end{pmatrix}_{MLx1}. \qquad (12)$$

Let **A** be the matrix

$$\mathbf{A} = \begin{pmatrix} f_1[1,1] & \cdots & f_T[1,1] \\ f_1[1,2] & \cdots & f_T[1,2] \\ \vdots & \ddots & \vdots \\ f_1[M,L] & \cdots & f_T[M,L] \end{pmatrix}_{MLxT} . \quad (13)$$

$$J(\boldsymbol{\alpha}) = (\mathbf{p} - \mathbf{A}\boldsymbol{\alpha})^T (\mathbf{p} - \mathbf{A}\boldsymbol{\alpha}) . \quad (14)$$

$\boldsymbol{\alpha}$ that minimizes (14) is given in (15)

$$\boldsymbol{\alpha} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{p} . \quad (15)$$

Evaluation of several sets of well defined basis functions performed. The evaluation results led to the conclusion that the most effective set, in terms of word error rate and minimum square error turned to be The Cosine Series Bivariate Polynomials:

$$\hat{P}_L[m,l] = \sum_{j=0}^{Q} \sum_{v=0}^{Q-j} \beta_{jv} \cos v \frac{l-1}{L} \cos j \frac{m-1}{M} . \quad (16)$$

Where Q is the bivariate polynomials series order. Eq. (16) can be represented in the form of (10):

$$T = \frac{Q^2}{2} + \frac{3Q}{2} + 1 , \quad (17)$$

$$g_{jv}(m,l) = \cos(v \frac{l-1}{L}) \cos(j \frac{m-1}{M}) . \quad (18)$$

The function basis definition is given

$$\begin{aligned} f_1(m,l) &= g_{0,0}(m,l) \\ f_2(m,l) &= g_{0,1}(m,l) \\ &\vdots \\ f_{Q+1}(m,l) &= g_{0,Q}(m,l) \\ f_{Q+2}(m,l) &= g_{1,0}(m,l) \\ &\vdots \\ f_{2Q+1}(m,l) &= g_{1,Q-1}(m,l) \\ &\vdots \\ f_T(m,l) &= g_{Q,0}(m,l) \end{aligned} \quad (19)$$

In addition, high frequency elements reduction improved the error rate. The resemblance of the 3D spectrogram to an image motivated the high frequency cut-off, which is commonly used by image compression techniques involved with DCT-based transforms. Figure 2 shows the scaled spectrogram in log scale, P_L vs. the approximated spectrogram as defined in (16). The original signal whose spectrogram plotted in figure (2) is a 100 [ms] frame taken from the word "zero". The short-term frame size was 20[ms]. The spectrogram was created with 9 short frames (10[ms] overlap), and 17 critical bands. L determines the spectrum range. It depends on the sampling frequency, the SNR, and the algorithm's application. For example, music compression utility would yield better quality as L rises towards the full audible frequency range. On the contrary, keyword spotting machine in a noisy environment should use such L that covers the low frequency range.

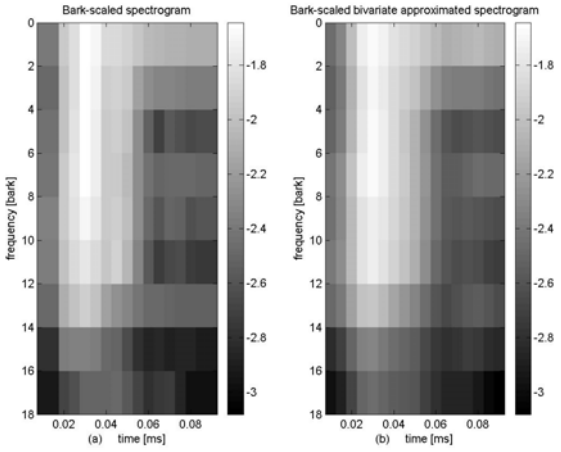

Figure 2: (a) Bark scaled psychoacoustic spectrogram. The coefficient vector size is 153 (b) Approximated bark scaled psychoacoustic spectrogram with the bivariate cosine basis (order 7). The coefficients vector size is 32.

4. RESULTS

4.1 Word Recognition

A training set consisting 224 occurrences of each digit by 224 speakers (i.e., a single occurrence of each digit per talker) was used. Half the talkers were male, half female.
A new set of 224 speakers (half male, half female) was used for testing.
The algorithm was tested on single word recognition and compared to the well-known algorithms MFCC, MFCC Δ, and MFCC $\Delta\Delta$.

Algorithm	Average Error Rate	Vector size	Frame size/overlap [ms]	Baud rate
MFCC	2.05%	13	25/10	1.105
MFCC Δ	0.71%	26	25/10	2.21
MFCC $\Delta\Delta$	0.36%	39	25/10	4.42
Described algorithm	.062%	32	110/-	0.29

Table 1: average digit recognition error rates for several recognizers. The baud rate is measured in [K coefficients/sec].

The database used is the "TIDIGIT", and the HMM platform is HTK. The experimental framework included four algorithms:

(i) MFCC: 13 coefficients (including the logged energy) created from 22 lifters.
(ii) MFCC Δ: 26 coefficients. 22 lifters, and the 0'th cepstral parameter.
(iii) MFCC $\Delta\Delta$: 39 coefficients. 22 lifters, and the 0'th cepstral parameter.

(iv) The supplied model: bivariate polynomials of order 7, with high frequency cut-off. 10 Short-term frames of 20[ms] with 10 [ms] overlap. 17 critical bands.

As it can be seen in Table 1 the MFCC based recognizer gives inferior performance than the suggested algorithm in a matter of error rate and baud rate. In addition the model's error rate equals to the MFCC Δ, but the MFCC Δ baud rate is about 6 times bigger.

4.2 Noisy Environment words recognition Evaluation

The experiment's objective is to evaluate the algorithm's immunity to noise. It included a database containing 50 male speakers for training, and other 50 male speakers for testing. The creation of the noisy environment, speech database, performed by adding sampled car noise to the original TIDIGIT in a given SNR. This procedure applied to the speech data base in different several Signal to Noise Ratios Ranged from 12[db] down to -3[db]. A typical car noise spectrum is presented in figure 3.

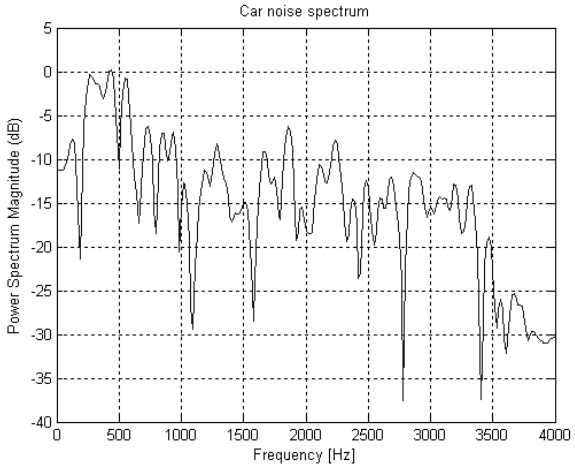

Figure 3: car noise spectrum

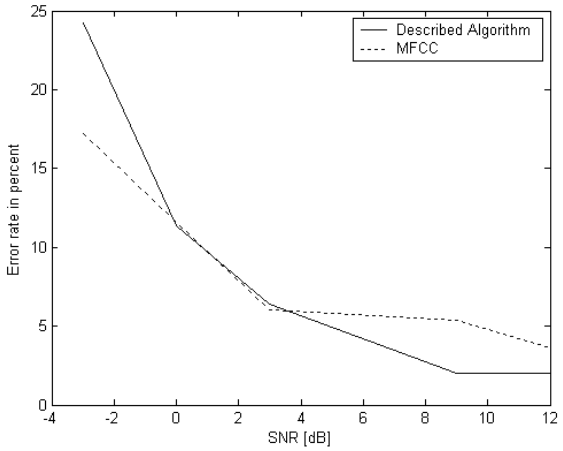

Figure 4: Error rate results Vs. SNR

Figure 4 shows the Average Word Recognition error rate as a function of the Signal to Noise Ratio (SNR).

5. CONCLUSION

In this paper, a new perceptual time-varying model of speech signal is presented.

This model was developed for ASR and compression applications. An important compression application property is baud-rate, and the ASR's important property is error-rate.

It can be seen that the proposed model achieves better recognition rate than the MFCC Δ. The proposed perceptual time varying model out perform in the baud rate point of view. The proposed model uses one third coefficients that is used in the case of MFCC, and one sixth that is used in MFCC Delta. Although the MFCC $\Delta\Delta$ perform better then the presented algorithm one should consider the baud rate ratio [4.42/0.29] Which enhances the advantage of the new model.

The proposed model seems more immune to noise than the MFCC at the positive SNR range. However it is more sensitive to additive noise at the negative SNR range. An interesting topic for further work might be improving the model's sensitivity to additive noise at low SNR.

ACKNOWLEDGMENT

The authors would like to thank Gabriel Zigelboim for his helpful comments on earlier versions of this paper.

REFERENCES

[1] H.Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, 87:1738-1752, 1990.

[2] F.Baumgarte, "Improved Audio Coding Using a Psychoacoustic Model Based on a Cochlear Filter Bank," *IEEE* Transactions on speech and audio processing, Vol. 10, No.7, October 2002.

[3] L.Cohen, "Time-Frequency Distributions – A Review," *Proceedings of the IEEE*, Vol. 77, No. 7, July 1989, p. 941-980.

[4] 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.