# Prediction of Speech Intelligibility Using a Neurogram Orthogonal Polynomial Measure (NOPM)

Nursadul Mamun, Wissam A. Jassim, and Muhammad S. A. Zilany

***Abstract*—Sensorineural hearing loss (SNHL) is an increasingly prevalent condition, resulting from damage to the inner ear and causing a reduction in speech intelligibility. This paper proposes a new speech intelligibility prediction metric, the neurogram orthogonal polynomial measure (NOPM). This metric applies orthogonal moments to the auditory neurogram to predict speech intelligibility for listeners with and without hearing loss. The model simulates the responses of auditory-nerve fibers to speech signals under quiet and noisy conditions. Neurograms were created using a physiologically based computational model of the auditory periphery. A well-known orthogonal polynomial measure, Krawtchouk moments, was applied to extract features from the auditory neurogram. The predicted intelligibility scores were compared to subjective results, and NOPM showed a good fit with the subjective scores for normal listeners and also for listeners with hearing loss. The proposed metric has a realistic and wider dynamic range than corresponding existing metrics, such as mean structural similarity index measure and neurogram similarity index measure, and the predicted scores are also well-separated as a function of hearing loss. The application of this metric could be extended for assessing hearing-aid and speech-enhancement algorithms.**

***Index Terms*—Auditory-nerve model, neurogram, orthogonal moment, sensorineural hearing loss, speech intelligibility.**

## I. Introduction

PERFORMING listening tests is an expensive, time consuming, and complicated operation, because it relies on subject's feedback and laboratory test conditions. However, subjective scores can be estimated by replacing the listeners with a model of the auditory system. Computational models of the normal and impaired auditory system are useful to detect, analyze, and segregate dynamic acoustic stimuli in complex environments. The availability of these models motivates the development of an objective measurement technique that predicts speech intelligibility [1]. To develop the metric, the full-reference method was used, in which neural responses to an original speech signal are compared with the neural responses to distorted speech.

The speech recognition performance of human subjects with normal hearing tends to decrease when speech is presented at high intensities (at sound pressure levels above 90 dB). In 1922, Fletcher first reported that the recognition performance of nonsense syllables presented in quiet environment decreases when the syllables are highly amplified [2].

In general, recognition performance of monosyllabic word drops significantly when speech level is increased beyond conversational speech level at a constant SNR, and the amount of deterioration may vary with the spectral content of the speech and masker [3]–[6]. Auditory masking occurs when perception of one speech signal is affected by the presence of another sound. The recognition performance of listeners with hearing loss also tends to decrease as a function of sensorineural hearing loss (SNHL) under quiet and noisy conditions [6]. Efforts have been made in the last few decades to develop a metric that can successfully assess speech intelligibility under various conditions.

Objective measures of speech intelligibility can be broadly classified into two categories. One approach is based on the properties of the acoustic signal and uses ad-hoc methods to address the effects of hearing loss and supra-threshold nonlinearities. The articulation index (AI) [7] and speech-transmission index (STI) [8] are common examples of this method. The other category of existing metrics uses computational models of the auditory system. Effects of hearing loss and supra-threshold nonlinearities are captured by the model of the auditory system itself. The spectro-temporal modulation index (STMI) [9], mean structural similarity index measure (MSSIM) [10], neurogram similarity index measure (NSIM) [11], and neural articulation index (NAI) [12] are examples of the second type of metric. In general, computational model-based metric appears to incur high computational complexity, because this type of metric requires simulating responses for an extensive number of neurons from the auditory system. Our proposed metric, NOPM is also based on a physiological computational model of the auditory system.

In 1947, the AI [7] was introduced as an objective measure, and it was the most widely used technique for predicting speech intelligibility of a transmission channel. The AI metric measured the signal-to-noise ratio (SNR) on a dB scale in a number of frequency bands covering the speech spectrum. The AI calculation scheme was further improved to increase its accessibility

Manuscript received January 07, 2014; revised May 23, 2014; accepted January 25, 2015. Date of publication February 06, 2015; date of current version March 16, 2015. This work was supported by the University of Malaya under High Impact Research Grant UM.C/625/1/HIR/152 (MSAZ). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wai-Yip Geoffrey Chan.

The authors are with the Department of Biomedical Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia (e-mail: nursad49@gmail.com; binaye2001@yahoo.com; msazilany@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2401513

and validity for use in the evaluation of most speech communication systems under a wide variety of noise and speech-distortion conditions [13], [14]. The AI has been modified to develop speech intelligibility index (SII) that takes into account the effects of auditory thresholds, higher presentation levels, self-masking, and upward spread of masking. However, it is well-known that time-domain distortions such as reverberation, echoes, and nonlinear distortions such as peak clipping have large negative effects on speech intelligibility of transmission channels [15]. Kates and Arehart [16] further extended the SII by using the signal-to-distortion ratio from the coherence instead of the SNR, and the resulting metric was able to capture effects of broadband noise and distortions for both normal-hearing and hearing-impaired listeners. The effect of fluctuating noise was included into the SII [17] by dividing the noisy speech into segments, computing the SII for each segment, and then averaging them across all segments. The STI [8], which is based on the modulation transfer function (MTF) of a transmission channel, takes into account the time-domain distortions by considering the envelope fluctuation rates encountered in running speech. Speech intelligibility predictions made by the AI and STI are related to the mean global transmission quality, which has been validated for phonetically balanced word tests, and their calculations are based on the average long-term speech spectrum.

The STMI [9] is a computational model-based metric that extracts spectro-temporal modulation information from the auditory neurogram. A neurogram is a two dimensional representation in which responses of neurons with different characteristic frequencies (CF) is displayed as a function of time. The STMI quantifies the degradation in the spectral and temporal modulations due to noise and can handle difficult and non-linear distortions such as phase-jitter (due to fluctuations of the power supply) and phase shifts, which cannot be handled by the STI. However, the STMI depends only on the slow-varying envelope and cannot explain speech intelligibility for various behavioral studies [18]. Few recent studies have reported that although envelope information from a few spectral bands is sufficient for speech intelligibility in quiet, TFS information is needed when speech is presented in a fluctuating or noisy background [19], [20]. In addition, the auditory periphery model used to compute STMI is a linear model and thus cannot handle non-linear effects of hearing loss observed at the level of auditory nerve (AN).

The auditory filters in the cochlea decompose the complex broadband sounds into a series of relatively narrowband signals, each of which can be considered as a slowly varying envelope (ENV) superimposed on a more rapid temporal fine structure (TFS). Although TFS information depends on phase locking to individual cycles of the stimulus waveform, both ENV and TFS information are represented in the timing of neural discharges [21]. Depending on the dominant fluctuation rates of speech signals, Rosen [22] separates the temporal features of speech into three primary categories: envelope, periodicity, and TFS. Envelope, which fluctuates at a rate (modulation frequency) between 2 and 50 Hz, conveys information of manner of articulation, voicing, vowel identity, and prosodic cues. Periodicity, which carries segmental information about voicing and manner and prosodic information relating to intonation and stress, fluctuates at rates between 50 and 500 Hz. TFS has dominant frequencies between 600 Hz to 10 kHz [12]. Smith *et al.* [18] observed that the envelope is most important for speech reception (speech intelligibility in quiet), and the TFS is important for pitch perception and sound localization. In general, recognition of English speech in quiet is dominated by the envelope, whereas recognition of melody is dominated by the TFS [23]. Pitch perception should also help convey prosody cues in speech and may enhance speech reception for tonal languages, such as Mandarin Chinese, where pitch is used to distinguish different words. Xu [23] observed that lexical-tone recognition depends on fine structure, not on envelope, when the number of frequency bands was between 4 and 16. Lorenzi *et al.* [24] reported that TFS cues convey more important phonetic information. Nie *et al.* [25] studied the importance of spectral and temporal cues in cochlear implant patients. Their results show the trade-off effect of spectral and temporal cues on speech intelligibility for implant users.

The dichotomy between the acoustic temporal ENV and TFS cues motivated the researchers to explore the relative role of ENV and TFS information in human speech perception. In this study, the contribution of neural ENV and TFS information (neurograms) to speech intelligibility is studied. Based on temporal resolution (bin width of 10 and 100 $\mu$s, which will subsequently be referred to as TFS and ENV, respectively), two types of neurogram were constructed from the output of the model for the auditory periphery. By applying image processing to the auditory neurogram, Hines [10] proposed the MSSIM to predict speech intelligibility for a range of sensorineural hearing losses (SNHLs). The MSSIM is a function of luminance, contrast, and structure of the neurogram, which was treated as an image.

Because there was not much effect of contrast on the MSSIM score, the NSIM was introduced by ignoring the effect of contrast. But both MSSIM and NSIM have relatively small dynamic ranges, meaning that the scores for two different extreme situations (flat 10 dB loss vs. profound hearing loss) do not vary substantially. For example, in Fig. 3 of Hines and colleagues [10] the dynamic range from flat 10 dB to profound hearing loss is approximately 0.5 to 0.1. Also there is a sharp decrease in the MSSIM and NSIM scores (from 1.0 to 0.5) from unimpaired to a flat 10 dB hearing loss (see Fig. 10 from Hines and Harte), which is unrealistic in that it does not agree with the results of behavioral studies [26], [27]. In general, if hearing thresholds of the listeners are within $10 - 15$ dB of the normal hearing, most behavioral studies treat them as normal listeners [3], [28]. So, the recognition score would be very close to that of normal hearing (very slow roll-off). On the other hand, it is expected that the recognition score for listeners with profound hearing loss ($> 60 - 70$ dB loss in octave frequencies ranging from 250 Hz to 8 kHz) would be very low at conversational speech level ($\sim 65$ dB SPL), because the signal would be inaudible for most frequencies.

In order to avoid the above-mentioned problems, a speech-intelligibility metric, the NOPM, is proposed in this study that employs orthogonal moments as a feature extractor from the auditory neurograms. Orthogonal moments have become widely used in various areas in image processing. Their application in-

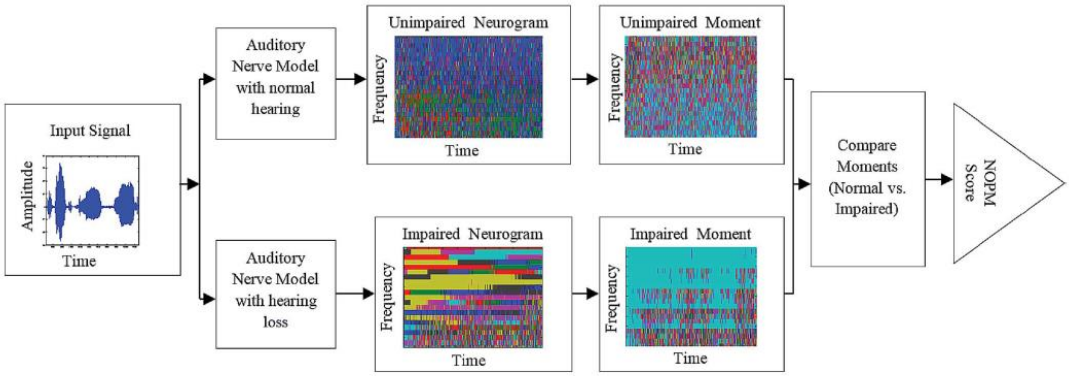


Fig. 1. Block diagram summarizing the methodology of NOPM. The speech signal was applied as an input to the models for the normal and impaired auditory systems, and the responses of the AN fibers with a wide range of characteristic frequencies were simulated to construct the neurograms. Neurograms were divided into blocks ($8 \times 8$) and orthogonal moments were applied on each block. Finally the features of the unimpaired and impaired moment coefficients were compared to provide an NOPM score.

cludes object identification or pattern recognition, image segmentation and edge detection, image compression, texture retrieval, and multi-resolution analysis [29], [30]. Due to their inherent properties such as information compactness, oscillating kernels and phase information of an image, orthogonal moments have successfully been employed in the aforementioned applications. Orthogonal moments have the ability to represent a signal using a limited number of coefficients without substantially compromising signal quality [31]. In addition to energy compaction, one of the most important properties of orthogonal moments is the ability to localize in space. So, selectively choosing a portion of the moment coefficients will allow reconstructing only a certain part of the original signal, which could contain the important perceptual feature of the acoustic signal (phoneme) to identify them. This motivated us to employ orthogonal moments in this study. However, DCT does not possess this property.

Motivated by the above-mentioned applications of orthogonal moments in various areas of image processing, a novel metric is proposed in this study to predict speech intelligibility under quiet and noisy conditions for listeners with normal hearing and also for listeners with a range of SNHLs. To our knowledge, there has been no previous study that uses orthogonal moments to extract features from auditory neurograms. Two types of orthogonal moments, the Tchebichef and Krawtchouk moments, are frequently used in the field of image processing. However, in this study only the Krawtchouk moments were employed to simulate the results, because the Tchebichef moments used as a feature extractor produced similar results.

This paper is organized as follows. Section II briefly introduces the phenomenological model of the auditory periphery employed in this study along with the components of the NOPM metric. Results of the NOPM metric using the TIMIT and NU6 database are described and compared with the results from the respective behavioral studies in Section III. Section IV discusses the important features of the metric with conclusions.

## II. Methods

The following sections briefly describe the computation of the proposed metric, the NOPM. In addition, existing computational model-based metrics, as well as the components of the proposed metric, will be briefly explained.

Fig. 1 briefly describes the procedure of the NOPM metric for listeners with hearing loss. The speech stimulus is applied to a computational model of the auditory periphery. Neurograms for normal and impaired auditory systems are then constructed, the neurogram features are extracted using orthogonal moment transform. Finally, the computed impaired and unimpaired moments are compared together to provide an NOPM score for a range of SNHLs. On the other hand, to predict speech intelligibility score under noisy conditions, both the clean and its corresponding noisy signals are applied to the model of the auditory system to construct clean (normal) and distorted (equivalent to impaired) neurograms.

### *A. Neurogram Orthogonal Polynomial Measure (NOPM)*

The neurogram orthogonal polynomial measure is an objective intelligibility measurement metric that uses orthogonal moments to quantify the change in signal information even for small changes in magnitude (pixel intensity) and phase (location) of the neurogram. It is well known that structural information significantly influences the quality of the image (neurogram). The local phase (discharge timing) of a neurogram contains more structural information than the magnitude, and any change in neurogram structure may lead to changes in the phase shifts. So, phase information is necessary to predict distortion precisely and the location of the distortion of a signal is as important as its magnitude [32]. Discrete orthogonal moments are good signal descriptors that can represent image information with minimum redundancy and are able to capture even small changes or differences of pixel intensity in an efficient manner and thus produce variations in the computed moment's values due to small changes of the pixel intensities.

*1) Auditory-Nerve Model:* The AN model [33] used in this study simulates the responses of the cochlea, inner hair cells (IHCs), outer hair cells (OHCs), and the IHC-AN synapse up to the responses of the AN fiber. The model successfully captures most of the nonlinearities such as compression, two-tone rate suppression, frequency selectivity, level-dependent rate, phase responses, and the shift in the best frequency (the frequency at which the fiber response is maximum) at higher levels observed at the level of the AN [33]–[38]. The model responses have been validated against a wide range of physiological recordings from AN fibers to simple (tone-like) and complex (speech-like) stimuli [39].

This study uses the AN model introduced by Zilany and colleagues [33], [40] to predict human speech-recognition performance. The schematic diagram of the current AN model is shown in Fig. 1 of Zilany *et al.* [37]. The model consists of some phenomenological functional components. The input instantaneous pressure waveform is passed to the middle ear, which is followed by the basilar membrane (BM) filter. The feed forward control path (which includes OHC functions) regulates the gain and bandwidth of the BM filter to account for the level-dependent properties of the cochlea. Basilar-membrane responses are passed through the IHC, which transduces the mechanical responses of the BM to an electrical potential. The IHC is modelled with a static nonlinearity followed by a low-pass filter. The spontaneous rate, adaptation properties, and rate-level function of the AN model are determined by the model of the IHC-AN synapse. The spike timings are provided by a non-homogenous Poisson process driven by the synapse output.

Noise-induced impairment in the cochlea causes damage to both the IHC and OHC stereocilia [39], [41], [42]. Damage to the OHC stereocilia causes both elevated threshold and broadened tuning of AN-fibers, whereas IHC stereocilia damage results only in the elevation of the tuning curve without any substantial effect on the bandwidth. The effects of the OHC and IHC status are incorporated in the model by introducing a scaling factor $\mathrm{C_{OHC}}(0 \leq \mathrm{C_{OHC}} \leq 1)$ to the control path output and $\mathrm{C_{IHC}}(0 \leq \mathrm{C_{IHC}} \leq 1)$ to the IHC transduction function, respectively. $\mathrm{C_{OHC}} = 1$ simulates the normal functioning and $\mathrm{C_{OHC}} = 0$ indicates complete impairment in the OHC. Similarly, the normal functioning of the IHC is represented by $\mathrm{C_{IHC}} = 1$, whereas $\mathrm{C_{IHC}} = 0$ corresponds to complete impairment in the IHC. These two scaling factors successfully capture the phenomena reported for damage to the OHC and IHC stereocilia [34], [37].

*2) Neurogram:* The neurogram is similar to a spectrogram, except that neural responses are simulated for a range of CFs as opposed to analyzing the acoustic waveform. In this study, neurograms were constructed by simulating the responses of the AN fibers to phonemes and words from the standard databases. The original speech token sampled at 16 kHz for TIMIT [43] and 44.1 kHz for NU#6 [44] was resampled at the rate required for the AN model (100 kHz) [33]. Responses of the AN for different presentation levels ranging from 60 to 99 dB SPL were simulated. In this study, responses of 32 AN fibers logarithmically spaced from 250 to 8000 Hz were simulated. Neural responses at each CF were simulated to the 50 repetitions of the same stimulus. To be consistent with the physiology, responses of three different types of AN fibers (high, medium, and low spontaneous rates) were simulated, and their responses were weighted according to the distribution of the spontaneous rates (high $= 0.6$, medium $= 0.2$, and low $= 0.2$ of total population) [39]. Two neurogram representations were initially created by averaging the neural responses of each CF with a bin (time window) size of 10 $\mu$s for TFS and 100 $\mu$s for ENV responses. The neural responses of each CF were then divided into frames using a Hamming window (50% overlap between adjacent frames) of length 32 samples for TFS and 128 samples for ENV [45], and the average value of each frame was calculated. The combination of binning to 100 $\mu$s and smoothing with the 128-sample accounted for spike synchronization to frequencies up to $\sim$ 160 Hz $[1/(100 \times 10^{-6} \times 128 \times 0.5)]$, whereas the binning at 10 $\mu$s and smoothing with the 32-sample Hamming window extended the range of included synchronization frequencies up to $\sim$ 6.25 kHz $[1/(10 \times 10^{-6} \times 32 \times 0.5)]$. So the ENV neurogram excludes spike timing information about the temporal fine structure, but the TFS neurogram includes it.

In order to simulate the responses for the hearing-impaired AN fibers, the model parameters for the inner hair cell ($\mathrm{C_{IHC}}$) and outer hair cell ($\mathrm{C_{OHC}}$) were varied from 1 to 0 according to the degree of hearing loss. For normal hearing, both of the parameters were set to 1, and the complete impairment in the IHC and OHC was implemented by setting both parameters to 0. The hearing loss profiles used in this study were flat 10 dB, flat 20 dB, mild, moderate, and profound hearing loss. The audiograms for the five profiles of hearing losses are shown in Fig. 2(A), and the corresponding scaling factors, $\mathrm{C_{IHC}}$ and $\mathrm{C_{OHC}}$, are shown in Fig. 2(B) and 2(C), respectively. The audiograms were taken from Dillon [46] to illustrate typical hearing losses at different frequencies. The hearing losses for 32 CFs were determined using interpolation, and the corresponding model parameters ($\mathrm{C_{IHC}}$ and $\mathrm{C_{OHC}}$) were evaluated. To compare the performance of the proposed method with the results from a behavioral study, specific hearing loss profiles (HI2 in [6]) were used and model parameters were estimated accordingly.

*3) Orthogonal Moments:* Orthogonal moments use orthogonal polynomials as a basis function and provide better feature representation capability than other real transforms (such as discrete cosine transform) and improved robustness to noise [47]. Discrete orthogonal moments can be used to transform signals (1-D or 2-D) from time or spatial domains to moment domain using a set of useful basis functions. Signals have more compact representations in the moment domain. Lower- and higher-order moments represent low- and high-frequency components of a signal, respectively [31]. Generally speaking, most of the signal energy in the moment domain is combined in the lower-order moments, whereas most of the noise energy is contained in the higher-order moments. However, in case of colored noise, both lower- and higher-order moments describe the noise [48], [49].

The inherent properties and mathematical modelling of a well-known set of orthogonal moments, the Discrete Krawtchouk Transform (DKT), which are formed based on the orthogonal polynomial kernels of the Krawtchouk polynomials, are briefly described as follows [47]. Orthogonal polynomials

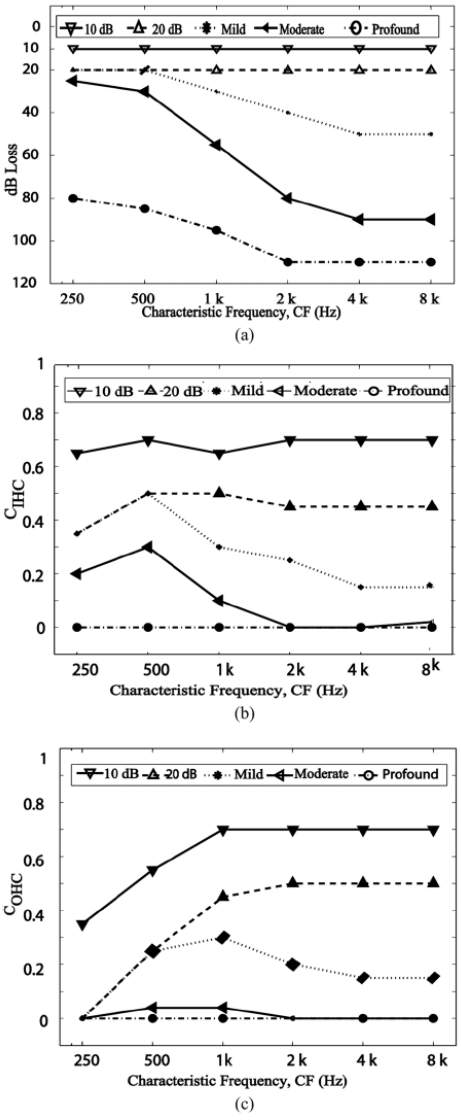


Fig. 2. Profiles for the hearing losses and the corresponding scaling parameters for the AN model. (A) Six profiles of hearing loss are shown in dB loss as a function of frequency, (B) Corresponding IHC parameter, and (C) Corresponding OHC parameter.

can be defined using a hyper-geometric series, ${}_rF_s$ which are given as follows:

$$ {}_rF_s(u_1,\ldots,u_r;v_1,\ldots,v_s;z) = \sum_{i=0}^{\infty} \frac{(u_1)_i(u_2)_i\ldots(u_r)_i z^i}{(v_1)_i(v_2)_i\ldots(v_s)_i(i!)} \tag{1} $$

where $(a)_i$ is a Pochhammer symbol and is defined as:

$$ (a)_i = a(a+1)\ldots(a+i-1) = \frac{\Gamma(a+i)}{\Gamma(a)} \tag{2} $$

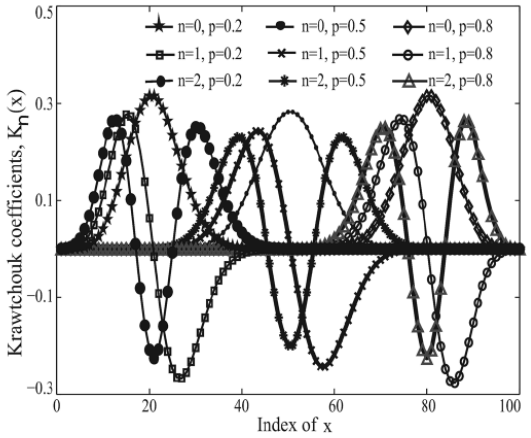


Fig. 3. The Krawtchouk polynomials plots for different values of polynomial order $n$, and parameter, $p$. $N = 100$.

It is not suitable to define orthogonal moments using general hypergeometric functions because when the moment order becomes larger the discrete orthogonal moments tend to exhibit numerical instabilities. Some computational aspects and recurrence algorithms can be used to calculate the polynomial coefficients. Moments with basis functions, including Krawtchouk polynomials, have better feature extraction capabilities than other moments [50]. Hence, orthogonal moments of Krawtchouk polynomials were used in this study as a feature extractor to quantify speech intelligibility.

*Krawtchouk Polynomials:* Orthogonal polynomials can be represented by a 2-D array with two parameters, $n$ (order of the polynomial) and $x$ (time or spatial index of the signal of length $N$). The $n$th-order normalized Krawtchouk polynomial, $K_n(x)$, for a signal of length $N$ is given by:

$$ K_n(x) = \sqrt{\frac{\binom{N-1}{x} p^x (1-p)^{N-1-x}}{(-1)^n \left(\frac{1-p}{p}\right)^n \frac{n!}{(-N+1)_n}}} \, {}_2F_1\left(-n,-x;-N+1;\frac{1}{p}\right) \tag{3} $$

where $p \in (0,1)$, $x = 0,1,\ldots,N-1$, and $n = 0,1,2,\ldots,N-1$. . The recurrence algorithms of $K_n(x)$ are given in [47], [51].

The value of $p$ controls the moment's localization on the region-of-interest (ROI). When $p = 0.5$, the ROI will be located in the middle of the signal frame. If $p < 0.5$ the ROI is shifted to the left, and for $p > 0.5$, the ROI is shifted to the right. Plots for the DKT matrix for few values of $n$ and $p$ are shown in Fig. 3 which illustrates the effect of the parameter $p$ on the position of ROI within the signal frame. Lower-order Krawtchouk polynomials can extract the low-frequency components of the different parts of the signal depending on parameter $p$. For example, the moment for $n = 0$, $p = 0.5$ extracts the low-frequency components of the middle part of the signal, whereas the moment for $n = 0$, $p = 0.2$ extracts the low-frequency components of the earlier part of the signal.

*Orthogonal Transformation:* The DKT for a block, $f(x,y)$, with a size of $N \times N$ extracted from the neurogram is defined as

$$\psi_{nm} = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} K_n(x) K_m(y) f(x,y) \tag{4}$$

where, $n, m = 0, 1, \ldots, N-1$. In matrix form:

$$\boldsymbol{\Psi} = \boldsymbol{K} * \boldsymbol{F} * \boldsymbol{K}^{\boldsymbol{T}}, \tag{5}$$

Here, $\boldsymbol{\Psi}(N \times N)$ represents the moment values of $f(x,y)$ in the transform domain. The operator $(*)$ refers to matrix multiplication, and $(\cdot)^{\mathrm{T}}$ refers to transpose of the matrix. $\boldsymbol{K}$ is the $N \times N$ dimensional polynomial matrices of Krawtchouk coefficients derived from Eq. (3), and $\boldsymbol{F}(N \times N)$ is a block extracted from the neurogram. The neurogram block can be reconstructed using the inverse transformation:

$$f(x,y) = \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} \psi_{nm} K_n(x) K_m(y)$$
$$x, y = 0, 1, 2 \ldots, N-1. \tag{6}$$

Also, Eq. (6) can be written in matrix form as

$$\boldsymbol{F} = \boldsymbol{K}^{\boldsymbol{T}} * \boldsymbol{\Psi} * \boldsymbol{K} \tag{7}$$

*4) Similarity Measure:* Cross-correlation is a measure of similarity of two signals or images and can be represented as a sliding dot product or inner-product. Correlations can be applied in pattern recognition, single particle analysis, and neurophysiology. In this study cross-correlation is used to compare two moment neurograms (neurograms transformed into moment domain) to provide quantitative results. The 2-D correlation coefficient between two images (A and B) can be calculated as follows:

$$\rho_{AB} = \frac{\sum_{i=1}^{u} \sum_{j=1}^{v} (A_{ij} - \mu_A)(B_{ij} - \mu_B)}{\sqrt{\sum_{i=1}^{u} \sum_{j=1}^{v} (A_{ij} - \mu_A)^2 \sum_{i=1}^{u} \sum_{j=1}^{v} (B_{ij} - \mu_B)^2}} \tag{8}$$

Here, $\mu_A$ and $\mu_B$ represent the mean value of image A and B, respectively.

*5) Test Corpora:* This section introduces the test corpora used to estimate the scores of NOPM metric. Two databases were used: TIMIT [43] and NU#6 [44]. TIMIT database was used to predict NOPM scores for listeners with hearing loss. TIMIT is a corpus of phonemically and lexically transcribed speech of American English speaker. The core portion of the test set contains 24 speakers, 2 male and 1 female from each dialect region. There are 7753 phoneme utterances in the core test set. The fifty seven distinct phoneme types in TIMIT were divided into six phoneme groups (Table I). These phoneme utterances were used in this experiment to quantify the speech intelligibility at different sound presentation levels for a range of SNHLs.

To compare the NOPM scores with the results from behavioral studies, NU#6 database was used for listeners with both normal hearing and hearing loss. The NOPM scores were predicted as a function of SPL and SNR (using speech-shaped noise) to compare with the scores from the behavioral studies of Studebaker *et al.* [6] and Dubno *et al.* [3]. NU#6 is a corpus

TABLE I
TIMIT PHONEME GROUPS FOR CORE TEST

| Phoneme group | No. in core set | Phonemes |
|---|---|---|
| Vowels | 2703 | /iy/, /ih/, /eh/, /ey/, /ae/, /aa/, /aw/, / ay/, /ah/, / ao/, /oy/, /ow/, / uh/, /uw/, /ux/, /er/, /ax/, /ix/, /axr/, /ax-h/ |
| Stops | 2363 | /b/, /d/, /g/, /p/, /t/, /k/, /dx/, /q/, /tcl/, /bcl/, /dcl/, /pcl/, /kcl/, /gcl/ |
| Affricates | 90 | /jh/, /ch/ |
| Fricatives | 1015 | /s/, /sh/, /z/, /zh/, /f/, /th/, /v/, /dh/ |
| Nasals | 670 | /m/, /n/, /ng/, /em/, /en/, /eng/, /nx/ |
| SV/glides | 1012 | /l/, /r/, /w/, /y/, /hh/, /hv/, /el/ |

of 200 monosyllabic words from a male speaker recorded by Auditec of St. Louis.

*6) Procedure:* The input to the model was the set of phonemes or words from the database; each utterance was up-sampled at a rate of 100 kHz required for the AN model. The high sampling rate was required in the model to ensure stability of the digital filters implemented for faithful replication of frequency responses of different stages (e.g., middle ear) in the peripheral auditory system [34], [37]. In this study, the output of the synapse model was used to construct ENV and TFS neurograms. In order to capture small changes in the neurogram, it was necessary to divide the neurogram into small blocks. The neurogram was first divided into blocks of $4 \times 4$, and then each block was appended by 50% overlap from the adjacent blocks on each side ($2 \times 4$ in top and bottom side, and $4 \times 2$ in left and right side), and thus the size of each resulting block became $8 \times 8$. The array of polynomials is generated according to the size of the block to be processed. The block size (N × N) was chosen such that the time corresponding to N does not exceed $\sim$ 32 ms for which a speech signal is quasi-stationary. The moment features were then computed for each block. From the transformed neurogram ($8 \times 8$), only the middle $4 \times 4$ part was chosen which reflects the maximum change in energy from normal to distorted condition. The correlation coefficient between normal and impaired (distorted) moment neurograms was computed to produce a NOPM score. Speech presented at 64 dB SPL in quiet for a normal listener was used as a reference to compute NOPM scores for all conditions.

### B. Existing Metrics

The NOPM results were compared to existing metrics, the mean structural similarity index and neurogram similarity index which are full-reference metrics to account for degradation in speech intelligibility due to hearing loss [10]. MSSIM is a function of luminance (l), contrast (c), and structure (s):

$$MSSIM(x,y) = f(l(x,y), c(x,y), s(x,y)) \tag{9}$$

where luminance, $l(x,y)$ is a measure that compares the mean values across the two signals (normal and distorted), the contrast, $c(x,y)$ measures the variance of two signals using the

relative standard deviation, the structure, $s(x,y)$ is an inner product of two vectors which is equivalent to the measurement of Pearson correlation coefficient.

MSSIM is defined as (See Wang *et al.* [52], Hines and Harte [10] for a full description)

$$MSSIM(x,y) = \left(\frac{2\mu_x\mu_y + C1}{\mu_x^2 + \mu_y^2 + C1}\right)^{\alpha} \cdot \left(\frac{2\sigma_x\sigma_y + C2}{\sigma_x^2 + \sigma_y^2 + C2}\right)^{\beta} \cdot \left(\frac{\sigma_{xy} + C3}{\sigma_x\sigma_y + C3}\right)^{\gamma} \tag{10}$$

Here, $\alpha$, $\beta$, $\gamma$ and are weights for luminance, contrast, and structure, respectively. $\mu_x$, $\mu_y$ are the mean and $\sigma_x$, $\sigma_y$ are the standard deviation of normal and impaired neurograms, respectively. In Hines, $\alpha = \beta = \gamma = 1$ and constant $C1 = 0.02$, $C2 = 0.03$, and $C3 = C2/2$ was employed. As contrast has no significant effect on the MSSIM, Hines *et al.* [11] proposed a new metric called NSIM, defined as follows:

$$NSIM(x,y) = \frac{2\mu_x\mu_y + C1}{\mu_x^2 + \mu_y^2 + C1} \cdot \frac{\sigma_{xy} + C2}{\sigma_x \cdot \sigma_y + C2} \tag{11}$$

In this study, the performance of the proposed metric, NOPM, has been compared to the performance of MSSIM and NSIM, because these are recent measures that employed the same phenomenological model of the auditory nerve [33] to construct auditory neurograms. In addition, both the proposed and the reference metrics used tools from image processing to extract features from the neurogram, and the performance of the metrics was evaluated for people with hearing loss.

## III. Results

This section provides simulation results of predicted speech intelligibility for listeners with normal hearing and hearing loss using the NOPM metric proposed in this paper. The estimated NOPM scores were then compared with the subjective scores from behavioral experiments. The effects of different parameters on the predicted score were considered, including the block size, ROI, SPL, and SNR. The estimated scores using existing measures including MSSIM and NSIM were compared to the scores predicted by NOPM. The time required to compute an NOPM score in response to a typical NU#6 word was approximately 28.2 secs (using a standard computer with a 64-bit operating system and a processor speed of 3.1 GHz).

### A. NOPM Score for Listeners with Hearing Loss

In this section, phonemes extracted from the TIMIT database were used to compute NOPM scores for a range of SNHL. Fig. 4 shows simulation results for six phoneme groups at 65 dB SPL. The NOPM scores are shown for different phoneme groups with their TFS and ENV responses.

The NOPM score at any point represents the correlation coefficients between unimpaired (normal) and impaired moments averaged across a phoneme group. In general, the NOPM score for both ENV and TFS progressively deteriorates as the degree of hearing loss becomes greater. The error bars show $\pm 1$ standard deviation from the mean. It is clear that the dynamic range of the NOPM score for the TFS neurogram is

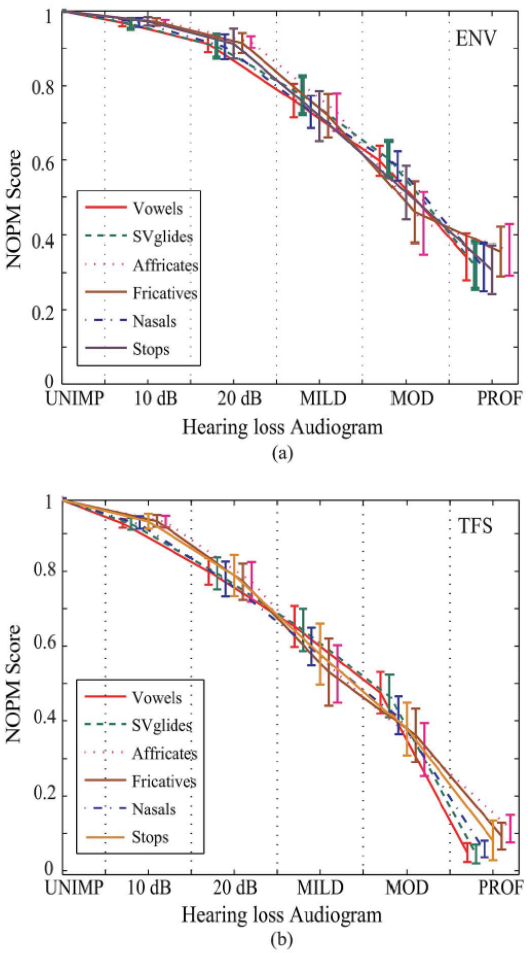


Fig. 4. NOPM scores for different phoneme groups using Krawtchouk moments. (A) NOPM scores using ENV response. (B) NOPM scores using TFS response. Results for all phoneme groups at 65 dB SPL are shown as a function of hearing loss. (A, B): NOPM scores using Krawtchouk moments for a window size of $8 \times 8$, constant $p = 0.9$, and moment order $N = 8$ with $\pm 1$ standard deviation.

higher (0.1-1.0) than the dynamic range of the envelope-based NOPM score (0.3-1.0). Among the six types of phonemes, the score based on TFS for vowels drops more compared to other phonemes when the hearing loss changes from mild to moderate. In general, it is interesting to note that the change in the NOPM score from unimpaired to 10-dB hearing loss is small compared to the noticeable decrease in the NOPM score for the other profiles of hearing loss. This result reflects the ability of the NOPM metric to predict closely the responses for the respective behavioral studies, as mentioned earlier. To examine the statistical significance, a pair-wise (two adjacent hearing-loss profile) t-test was applied to the NOPM scores for all phoneme groups. The level of significance (p) was found to be less than 0.01 ($p < 0.01$).

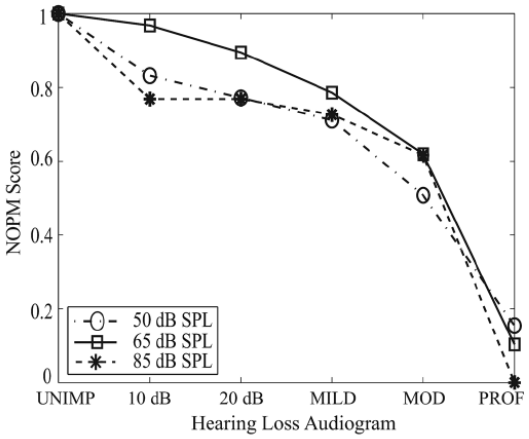


Fig. 5. Effects of SPL on NOPM score. Results are shown for Vowels using TFS response at 50, 65, and 85 dB SPL for $p = 0.9$, a window size of W1 $= 8 \times 8$, and moment order of $N = 8$.

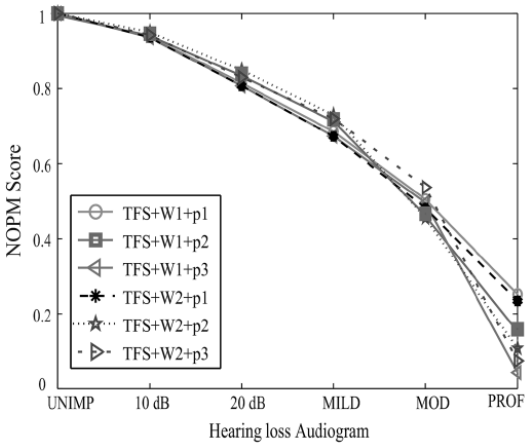


Fig. 6. Effects of the parameter and block size, W, on NOPM scores. Results for vowels using TFS responses at 65 dB SPL for $p1 = 0.2$, $p2 = 0.5$, and $p3 = 0.9$, and window size of $W1 = 8 \times 8 (N = 8)$ and $W2 = 16 \times 16 (N = 16)$, are shown. Here order of moment is equal to block size.

### B. *Effects of SPL on NOPM Score*

In general, speech intelligibility increases for people with hearing loss (moderate to severe) when the SPL is increased from lower sound levels to a conversational speech level ($\sim 65 - 70$ dB SPL). However, when the speech level is increased beyond the conversational speech level, the recognition performance under quiet condition declines for people with hearing loss [53]. This phenomenon is referred to as the roll-over effect in the literature. In order to quantify the effect of SPL on the proposed metric, the NOPM scores were estimated for vowel phoneme groups at 50, 65 and 85 dB SPL in Fig. 5. In general, the NOPM score at a constant SPL declines as a function of SNHL. The result also shows that the NOPM score increases when SPL increases from 50 dB to 65 dB and decreases when speech is more highly amplified (85 dB SPL).

This decline at a high level could be attributed to the loss of synchrony capture by higher formants and spread of excitation, which has been successfully captured by the AN model employed to construct the auditory neurograms [54]. In response to a vowel at higher sound levels, AN responses show the loss of synchrony capture by the second formant whereas synchrony to the first formant increases [33], [54].

### C. *Effects of Moment Parameters*

The effects of block size, W, and parameter, $p$, on the NOPM metric are illustrated in Fig. 6. The NOPM scores for two different block sizes ($W_1 = 8 \times 8$, $W_2 = 16 \times 16$) as well as parameter $p$ (0.2, 0.5, and 0.9) for representative vowel phonemes were simulated. The NOPM score based on the ENV responses was sensitive to both block size and parameter $p$ (not shown in figure), whereas scores based on the TFS responses were highly sensitive to parameter $p$ and less sensitive to parameter $W$, particularly for moderate and profound hearing loss. As the value of parameter $p$ increased, the contrast among NOPM scores also increased, especially for ENV responses. A detailed analysis of plots from the other phoneme groups revealed similar behavior (results not shown). In addition, it was found that the correlation coefficient between the original neurogram and the reconstructed neurogram using middle order moments with $p = 0.9$ was higher compared to using middle order moments with any lower values of the parameter $p$. This suggests that middle order moments with $p = 0.9$ captured relatively more information from the original neurogram, and thus the result for all subsequent figures is reported for $p = 0.9$.

### D. *Effects of SNR on the Predicted Score for Listeners with Normal Hearing*

Fig. 7 shows the estimated mean scores of NOPM, SII[1] [55], and NSIM as a function of SNR. The scores were predicted for NU#6 words using the AN model for a normal listener, and the SNRs were varied from $-20$ to $+60$ dB in steps of 10 dB. Additive white Gaussian noise (AWGN) and speech-shaped noise were used as maskers, and the words were presented at 65 dB SPL. The NOPM scores were computed using the TFS responses. In general, the scores progressively decreased as more and more noise was added to the signal.

This negative effect of increasing noise levels on the objective scores is consistent with the results from the relevant behavioral studies [3]. Studebaker *et al.* [6] also reported that the effective dynamic range of speech may be considerably larger than the commonly assumed value of 30 dB (e.g., in AI calculation). From this figure, it is obvious that the proposed metric (NOPM) and NSIM successfully captured this phenomenon.

### E. *Comparison with Existing Metric*

Fig. 8 compares NOPM scores with MSSIM and NSIM scores for the stops phoneme group using TFS responses. In contrast to the MSSIM and NSIM score, the NOPM score

[1]The Matlab code given in [55] is used to compute the SII scores for NU#6 words

Link to Full-Text Articles :

***http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7035038&tag=1***