INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

IJARIIT

ISSN: 2454-132X

Impact factor: 4.295

(Volume 4, Issue 4)

Available online at: www.ijariit.com

ESTOI for predicting the Intelligibility of speech

Nitesh Kumar
niteshkumar277@gmail.com
Bangalore Institute of Technology, Bengaluru, Karnataka

Karunavathi R K
rkk_dcj@yahoo.com
Bangalore Institute of Technology, Bengaluru, Karnataka

ABSTRACT

Intelligibility listening tests are necessary during development and evaluation of speech processing algorithms, despite the fact that they are expensive and time-consuming. The proposed scheme uses a monaural intelligibility prediction algorithm, which has the potential of replacing some of the listening tests. The proposed algorithm shows similarities to the Short-Time Objective Intelligibility (STOI) algorithm but works for a larger range of input signals. In contrast to STOI, Extended STOI (ESTOI) does not assume mutual independence between frequency bands. ESTOI also incorporates spectral correlation by comparing complete 400-ms length spectrograms of the noisy/processed speech and the clean speech signals. As a consequence, ESTOI is also able to accurately predict the intelligibility of speech contaminated by temporally highly modulated noise sources in addition to noisy signals processed with time-frequency weighting. We show that ESTOI can be interpreted in terms of an orthogonal decomposition of short-time spectrograms into intelligibility subspaces, i.e., a ranking of spectrogram features according to their importance to intelligibility.

Keywords— *Articulation index, Speech intelligibility index, Speech transmission index, Extended SII, Short time objective intelligibility, Extended STOI*

1. INTRODUCTION

When developing speech communication systems for human receivers, listening tests play a significant role each for observation progress within the development section and for validating the performance of the final system. Often, listening tests are used to quantify aspects of speech quality and intelligibility. Though listening tests represent the only tool offered for measure ground-truth end-user impact, they're time-consuming, they'll need special auditive stimuli information and equipment, and they need the availability of a bunch of typical end-users. For these reasons, listening tests are expensive and might generally not be used repeatedly throughout the development section of an auditory communication system. Hence, cheaper alternatives or supplements area unit of interest.

2. BACKGROUND

In this paper, we have a tendency to specialize in intrusive, monaural intelligibility prediction models, i.e., algorithms that – instead of conducting AN actual listening test – predict the end result of the listening test based on the auditory stimuli of the test. historically, 2 lines of analysis serve as the inspiration for existing intelligibility prediction models: i) the Articulation Index (AI) by French and steinberg, that was later refined and standardized as the speech intelligibility Index (SII), and ii) the Speech Transmission Index (STI) bySteeneken and Houtgast.

AI and SII were developed with simple linear signal degradations, e.g., additive noise, in mind. To estimate intelligibility, the methods divide the signal under analysis into frequency sub-bands and assume that every sub-band contributes independently to understandability. The contribution of a sub-band is found by estimating the long-term speech and noise power within the sub-band to reach the long-term sub-band signal-to-noise ratio (SNR). Then, sub-band SNRs are restricted to the range from -15 to +15 decibel, normalized to a value between zero and one, and combined as a perceptually weighted average.

3. PROPOSED SYSTEM

ESTOI is a function of the noisy/processed signal x(n) and clean speech signal s(n). First, the signals are passed through a one-third octave filter bank, and the temporal envelopes of each sub band signal are extracted. The resulting clean and noisy/processed short-time envelope spectrograms are time- and frequency-normalized before the “distance” between them is computed, resulting in intermediate, short-time intelligibility indices dm. Finally, the intermediate indices are averaged to form the final intelligibility index d. The overall structure of the proposed intelligibility predictor, ESTOI, is outlined in Figure 1 ESTOI is a function of the noisy/processed signal under study x(n), and the clean, undistorted speech signal s(n). The goal of ESTOI is to produce a scalar output d, which is monotonically related to the intelligibility of x(n).

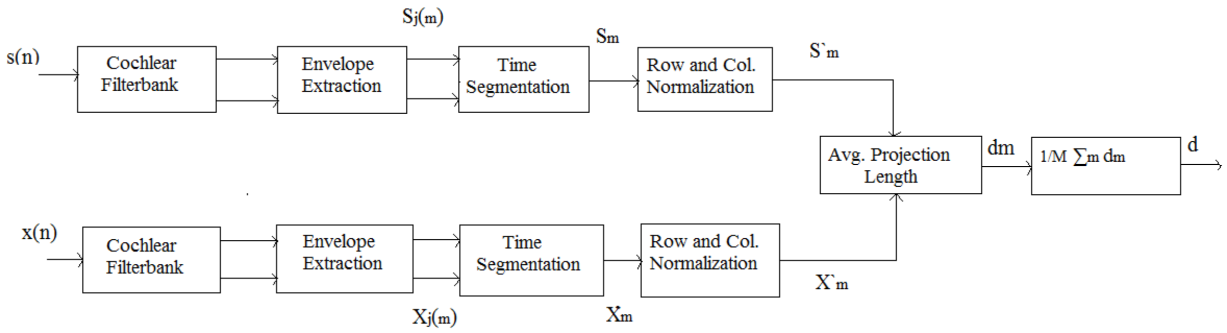


Fig. 1: The proposed intelligibility predictor, ESTOI

3.1 Time-Frequency Normalized Spectrograms

Give us a chance to accept that s(n) and x(n) are splendidly time-aligned, furthermore, that districts, where s(n) demonstrates no discourse movement (e.g., delays between sentences), have been expelled from the two signs. In the accompanying, we show expressions related to the clean signal s(n); comparative expressions hold for the loud/processed signal x(n). Give S(k,m) a chance to indicate the short-time Fourier transform (STFT) of s(n), that is

$$S(k,m) = \sum_{n=0}^{N'-1} s(mD+n)w(n)e^{-j2\pi kn/N'},$$

Where, k and m denote the frequency bin index and the frame index, respectively, and D and N′ denote the frameshift in samples and FFT order, respectively. Finally, w(n) is an analysis window.

To model crudely the signal transduction in the cochlear inner hair cells, a one-third octave band analysis is approximated by summing STFT coefficient energies,

$$S_j(m) = \sqrt{\sum_{k \in CB_j} |S(k,m)|^2}, \qquad j = 1, \ldots, J,$$

Where, j is the one-third octave band index, CB_j denotes the index set of STFT coefficients related to the jth one-third octave frequency band, and J denotes the number of subbands.

Let us collect spectral values S_j(m) for each frequency band j = 1,.....J, and across a time segment of N spectral samples, and arrange these in a short-time spectrogram matrix

$$S_m = \begin{bmatrix} S_1(m-N+1) & \cdots & S_1(m) \\ \vdots & & \vdots \\ S_J(m-N+1) & \cdots & S_J(m) \end{bmatrix}$$

Hence, the jth row of S_m represents the temporal envelope of the signal in sub band j. Typical parameter choices are J = 15 and N = 30 (corresponding to 384 ms) [14]. The noisy/processed short-time spectrogram matrix Xm is defined analogously.
ESTOI operates on mean- and variance- normalized rows and columns of Sm (and MX) as follows. Let

$$S_j, \mathrm{m} = [S_j(\mathrm{m}-\mathrm{N}+1)S_j(\mathrm{m}-\mathrm{N}+2)\ldots S_j(\mathrm{m})]^T$$

denote the jth row of the spectrogram matrix Sm. The jth mean- and variance- a normalized row of S_m is given by

$$\bar{s}_{j,m} = \frac{1}{\|(s_{j,m} - \mu_{s_{j,m}})\|}\left(s_{j,m} - \mu_{s_{j,m}}\underline{1}\right), \qquad (1)$$

Where, $\|y\| = \sqrt{(y^t y)}$ is the vector 2-norm, 1 is an all-one vector, and $\mu s_{j,m}$ is the sample mean given by

$$\mu_{s_{j,m}} = \frac{1}{N}\sum_{m'=0}^{N-1} S_j(m-m'). \qquad (2)$$

Note that the example mean and variance of the components in the vector $s^{-}_{j,m}$ is zero and one, separately. The mean and variance standardized rows $x^{-}_{j,m}$ of the noisy/processed signal are characterized comparably.

As specified, this line standardization technique is like the one utilized as a part of STOI. In particular, STOI utilizes a transitional worldly relationship coefficient for the jth sub-band in the mth time section, which can be communicated as the internal result of standardized vectors,

$$\bar{s}^T_{j,m}\bar{x}_{j,m}. \qquad (3)$$

However, as mentioned, in *ESTOI* we do not use Eq. (3) directly, but introduce a spectral normalization as follows. Let us first define the row-normalized spectrogram matrix

$$\bar{S}_m = \begin{bmatrix} \bar{s}^T_{1,m} \\ \vdots \\ \bar{s}^T_{J,m} \end{bmatrix}.$$

Then, let $s^{\vee}_{n,m}$ denote the mean- and variance- normalized nth column, n = 1, . . . ,N of m atrix s^{-}_{m}, where the normalization is carried out analogously to Eqs. (1) and (2).We finally define the row- and column-normalized matrix $\check{S}_m$ as

$$\check{S}_m = [\check{s}_{1,m} \cdots \check{s}_{N,m}] \,.$$

3.2 Intelligibility index

The row- and column- normalized matrices $\check{S}_m$ and $X^{\vee}_{m}$ fill in as the reason for the proposed understandability indicator. Specifically, we characterize a moderate understandability list, elated to time portion m, basically as

$$d_m = \frac{1}{N} \sum_{n=1}^{N} \check{s}^T_{n,m} \check{x}_{n,m}. \qquad (4)$$

Since $s^{\vee}_{n,m}$ and $x^{\vee}_{n,m}$, n = 1, . . . ,N are unit-standard vectors, each term in the aggregate might be perceived as the (marked) length of the orthogonal projection of the noisy/prepared vector $x^{\vee}_{n,m}$ onto the spotless vector $s^{\vee}_{n,m}$ or the other way around. It takes after that $-1 \le s^{\vee T}_{n,m} x^{\vee}_{n,m}$ ˇxn, m ≤ 1. So also, d_m might be deciphered as the (marked) length of these projections, arrived at the midpoint of crosswise over time inside a period portion. In low-commotion circumstances where $x^{\vee}_{n,m} \approx s^{\vee}_{n,m}$, at that point d_m will be near its most extreme normal projection length of 1, though if the components of and $s^{\vee}_{n,m}$ are uncorrelated, at that point dm ≈ 0, i.e., the vectors are around orthogonal. Additionally, from the meanings of $s^{\vee}_{n,m}$ and $x^{\vee}_{n,m}$, d_m might be translated just as test relationship coefficients of the sections of $S^{-}_{m}$and X^{-}_{m} (i.e., spectra which have been standardized by their sub band envelopes), arrived at the midpoint of over the N outlines inside a fragment. For straightforwardness, the understandability record identified with the whole loud/handled flag of intrigue is then characterized as the transient normal of the halfway clarity lists,

$$d = \frac{1}{M} \sum_{m=1}^{M} d_m, \qquad (5)$$

Where M is the number of time segments in the signal of interest. Since $-1 \le d_m \le 1$, it follows that −1 ≤ d ≤ 1.

3.3 Implementation

ESTOI works at an examining recurrence of 10 kHz to guarantee that the recurrence area applicable for discourse comprehensibility is secured [15]; all signs are resampled to this recurrence before applying the strategy. At that point, signals are isolated into casings of 256 examples, utilizing an edge move of D = 128, the edges are windowed with a Hann window, and a FFT of request N′ = 512 is connected. Before registering the clarity record, outlines with no discourse content are disposed of. These are recognized as the casings of the reference discourse flag s(n) with vitality under 40 dB than the flag outline with most extreme vitality. DFT coefficients of discourse dynamic edges are assembled into J = 15 33% octave groups, with focus frequencies of 150 Hz and around 4.3 kHz, for the most minimal and most astounding band, separately. At last, time fragments of length N = 30 (relating to 384 ms) are utilized.

4. SOFTWARE USED TO DESIGN THE PROPOSED SYSTEM

MATLAB (framework research facility) is a numerical handling condition and fourth-period programming dialect. Made by Math Works, MATLAB licenses interfacing with programs written in various dialects. Disregarding the way that MATLAB is proposed basically for numerical handling, graphical client interfacing, and furthermore to process video/picture.

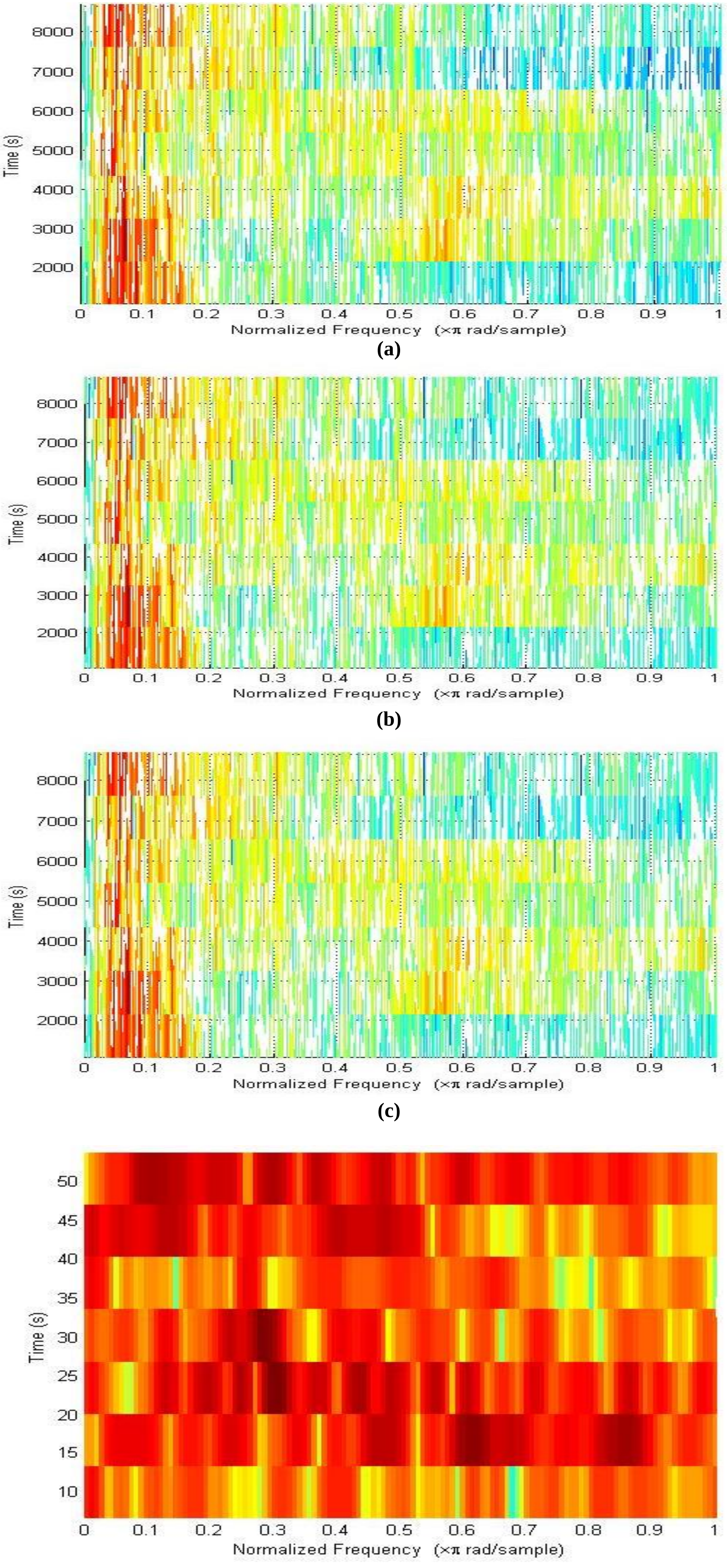
4.1 Strengths of MATLAB

- MATLAB is generally simple to learn.
- MATLAB code is improved to be moderately brisk when performing network tasks.
- MATLAB may act like a number cruncher or as a programming dialect.
- MATLAB is translated, blunders are less demanding to settle.
- Although essentially procedural, MATLAB has some question arranged components.

5. RESULTS AND DISCUSSION

In this area, we show various comprehensibility attentive tests for assessing the intended technique. Moreover, we contemplate the execution as an element of the section length N & test signal time. At long last, we contrast the execution of ESTOI with a scope of existing speech intelligibility indicators.

Figure 2 shows Short-time spectrograms for clean discourse time section (left segment) and uproarious time fragment (right segment) for added substance, discourse formed, sinusoidal adequacy regulated Gaussian clamor (adjustment recurrence of 5 Hz, SNR = - 10 dB). a), b) Time space sections. c), d) DFT short time spectrograms |S(k,m)|, |X(k,m)| (dB scale), registered by applying & N′ = 512 guide FFT toward zero padded, Hann windowed, time-area casings of 256 examples (25.6 ms) with a cover of D = 128 examples. e), f) third-arrange octave filter bank spectrograms s_m, x_m (dB scale). g), h) spectrograms with mean-and fluctuation standardized columns s^{-}_{m}, x^{-}_{m} (direct scale). I), j) spectrograms with mean-and difference standardized lines and sections, $\check{S}_m$, $x^{\vee}_{m}$(direct scale).

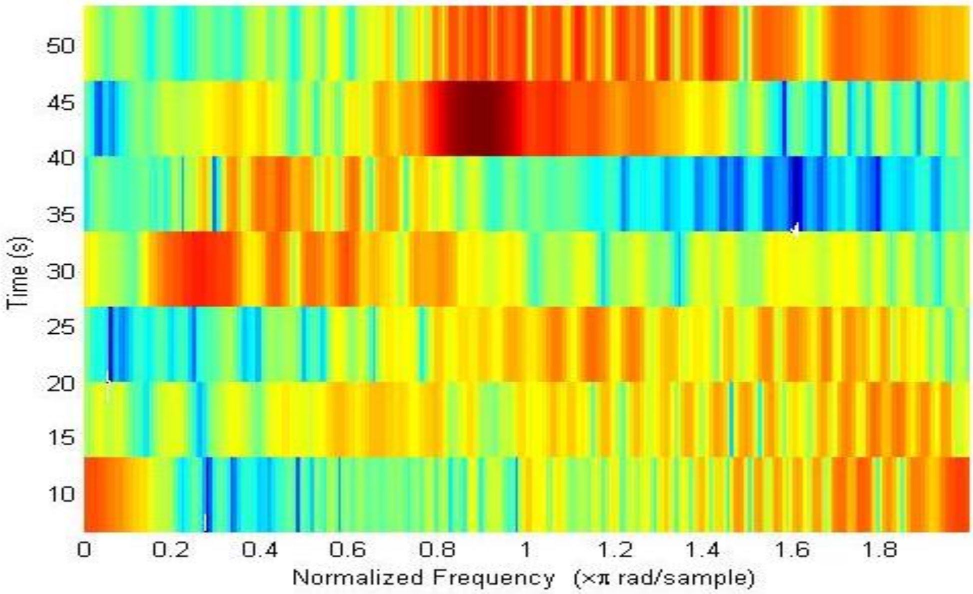

8000
7000
6000
5000
4000
3000
2000
Time (s)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Normalized Frequency (×π rad/sample)
50
45
40
35
30
25
20
15
10

(a)

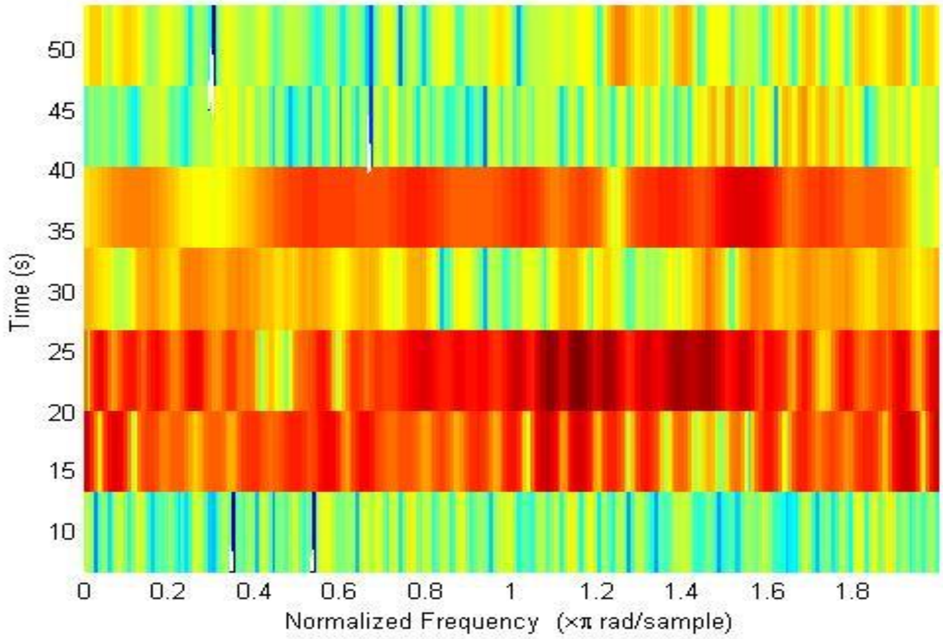
(b)

(c)

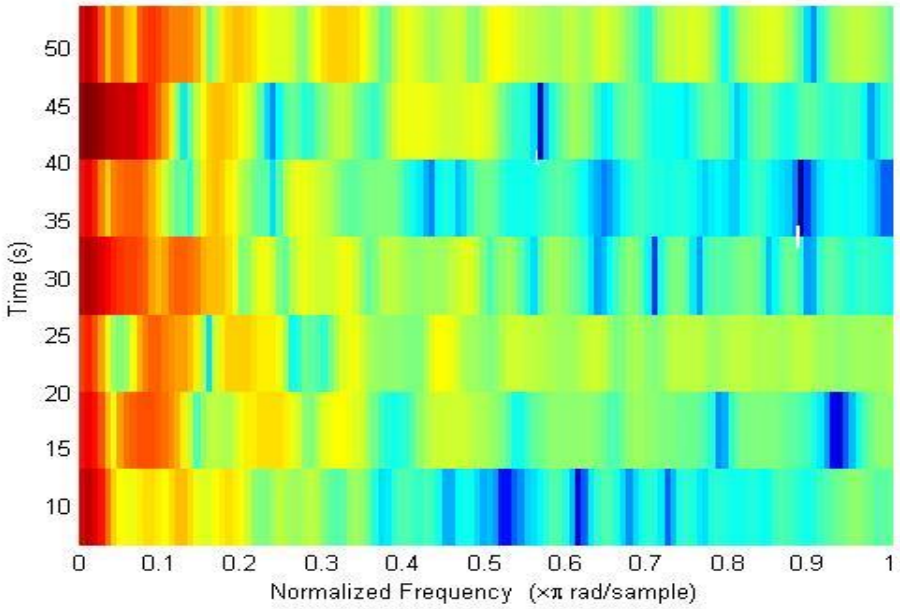
(d)


50
45
40
35
30
25
20
15
10
Time (s)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Normalized Frequency (×π rad/sample)

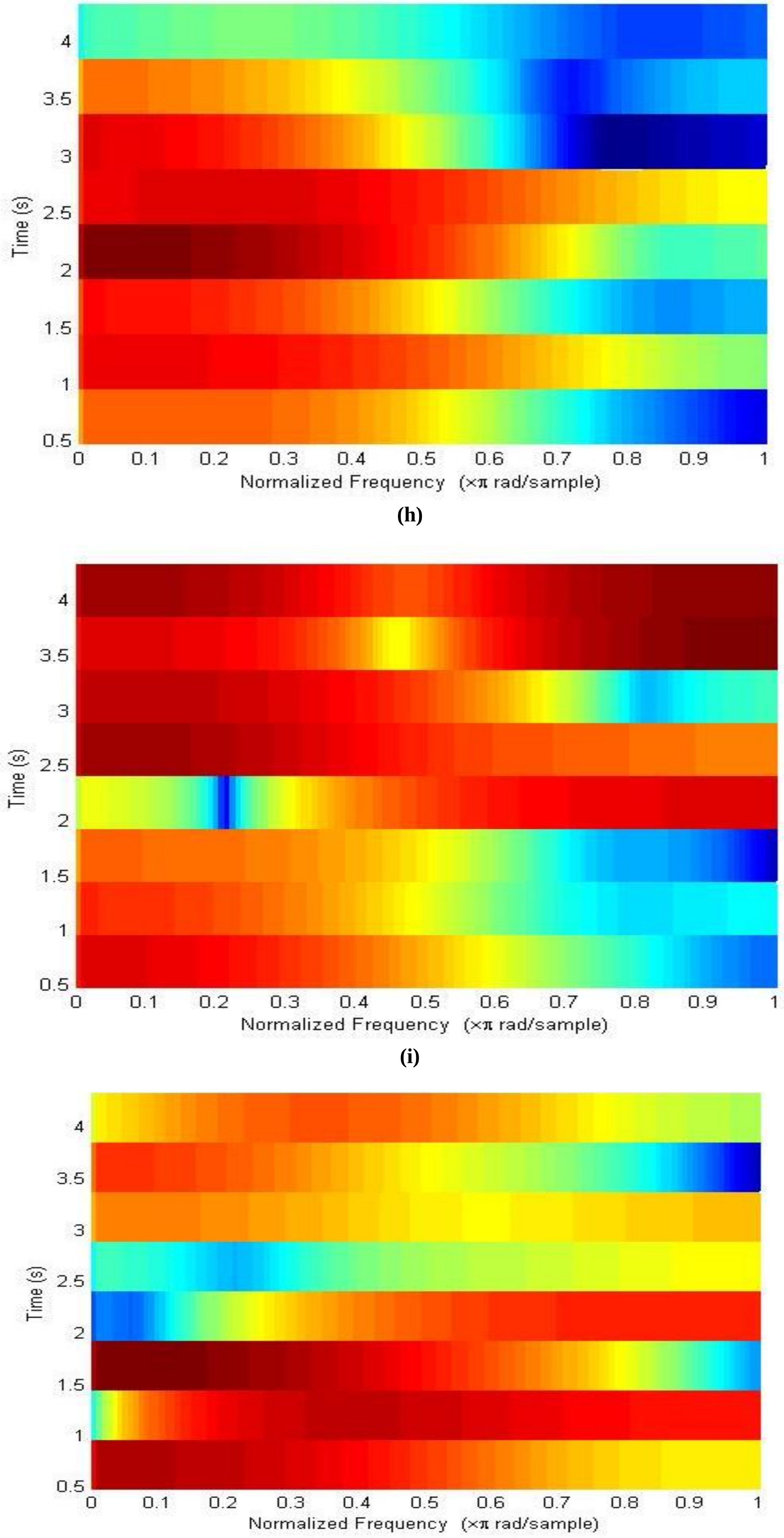
(e)


50
45
40
35
30
25
20
15
10
Time (s)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Normalized Frequency (×π rad/sample)

(f)


50
45
40
35
30
25
20
15
10
Time (s)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Normalized Frequency (×π rad/sample)

(g)



(h)

(i)

(j)

Fig. 2: Results

6. CONCLUSION

We presented an algorithm for monaural, intrusive intelligibility prediction. The algorithm estimates the average intelligibility of the potentially non-linearly processed signals, across a group of normal-hearing distortions. The proposed algorithm, which is called ESTOI, may be interpreted in terms of an orthogonal decomposition of normalized short time spectrograms into "intelligibility subspaces". These subspaces are ranked according to their importance with respect to intelligibility. This decomposition indicates, that the proposed algorithm forms spectro-temporal modulation patterns. The proposed intelligibility predictor has only one frame parameter, the segment length 'N' across which short time spectrograms are computed. We show, that the performance is fairly insensitive to the exact choice of this parameter & durations in the range of 256-640ms lead to best performance. We study the execution of ESTOI in predicting the result of five diverse understandability listening tests: two with temporally very regulated added additive noise sources, one with all the more reasonably balanced, added substance commotion sources, and two with boisterous signs prepared by perfect time-recurrence covering and single-channel non-direct commotion lessening calculations, separately. Contrasted with a scope of existing discourse comprehensibility expectation calculations, ESTOI does good overall hearing tests.

The present investigation has concentrated on discourse understandability expectation execution inside informational collections that each contains signals with comparable twists or handling composes. It is a point for future research to think about the execution of the proposed clarity indicator crosswise over informational indexes with various mutilation and handling composes. Contrasted with the present examination, this would require conduction of bigger understandability tests, where these diverse bending or handling writes are incorporated into a similar listening test.

7. REFERENCES

[1] R. McCreery, R. Ito, M. Stratford, D. Lewis, B. Hoover, and P. G. Stelmachowicz. Performance-intensity functions for normal-hearing adults and children using computer-aided speech perception assessment. Ear and Hearing, 31(1):95–101, 2010.

[2] Amyn M. Amlani, Jerry L. Punch, Teresa Y.C. Ching. "Methods and Applications of the Audibility Index in Hearing Aid Selection and Fitting".

[3] Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain, https://doi.org/10.1121/1.4964505.

[4] B. C. J. Moore. Cochlear Hearing Loss - Physiological, Psychological and Technical Issues. John Wiley and Sons, 2 edition, 2007.

[5] B. C. J. Moore. Cochlear Hearing Loss - Physiological, Psychological and Technical Issues. John Wiley and Sons, 2 edition, 2007.

[6] M. Zilany and I. Bruce. Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. The Journal of the Acoustical Society of America, 120(3):1446–1466, Sept 2006.

[7] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Amer., vol. 117, no. 4, pp. 2181–2192, 2005.

[8] T. Jurgens and T. Brand. Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. The Journal of the Acoustical Society of America, 126(5):2635–2648, 2009.

[9] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am., vol. 67, pp. 318–326, 1980.