INTERSPEECH 2020
OCTOBER 25-29/ SHANGHAI, CHINA
SHANGHAI INTERNATIONAL CONVENTION CENTER

Tutorial A-4-1

25 October 2020

Intelligibility Evaluation and Speech Enhancement based on Deep Learning

Fei Chen (Felix)

Department of Electrical and Electronic Engineering

Southern University of Science and Technology, Shenzhen, China

https://eee.sustech.edu.cn/feichen, fchen@sustech.edu.cn

CHEN, Fei (Felix)

Professor
Department of Electrical and Electronic Engineering
Southern University of Science and Technology (SUSTech), Shenzhen

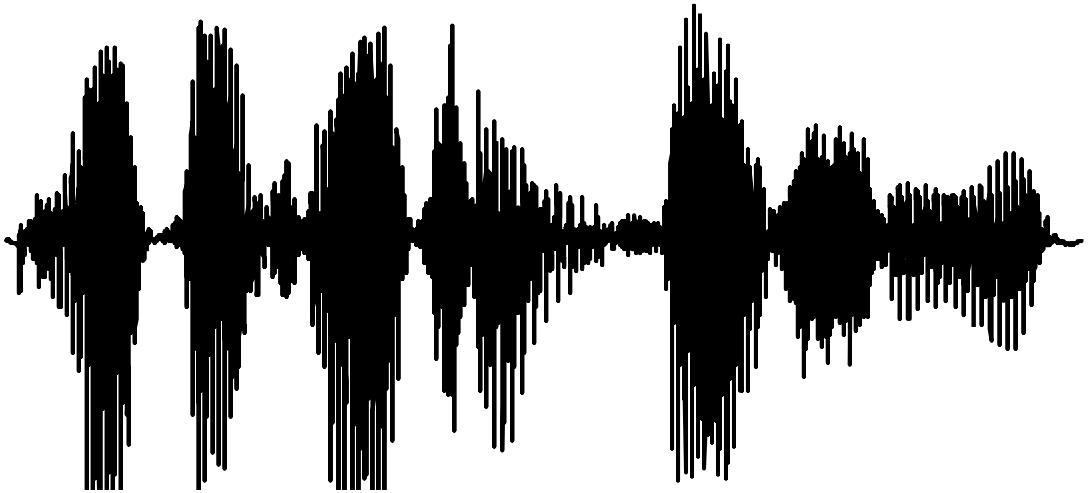
- 2020 – Present Professor, Southern University of Science and Technology
- 2014 – 2020 Associate Professor, Southern University of Science and Technology
- 2012 – 2014 Research Assistant Professor, The University of Hong Kong
- 2009 – 2011 Post-doctor Research Fellow, The University of Texas at Dallas [Dr. P. C. Loizou]
- 2001 – 2005 Ph.D., The Chinese University of Hong Kong

- Research areas: **speech perception, speech enhancement, intelligibility and quality modeling**.
  - Associate Editor/Editorial board member of <Biomedical Signal Processing and Control>, <Physiological Measurement> and <Frontier in Psychology>.
  - Published over 80 Journal papers in JASA, Ear and Hearing, JSLHR, Hearing Research, Speech Communication, etc.
  - Guest Editor for a Special Issue "Speech perception of Chinese-speaking Cochlear Implantees" at International Journal of Audiology.
  - Organized a Special Session "Signal Processing for Assistive Hearing Devices" at ICASSP 2015 in Brisbane, a Special Session "Biomedical Signal Processing in Assistive Hearing Systems" at ISIC 2014 in Singapore, and a Special Session of "Speech Recognition at Adverse Acoustic Conditions" at ISCSLP 2016 in Tianjin, a Special Session "Speech and Communication" at Inter-Noise 2017 in Hong Kong, etc.

Table of contents

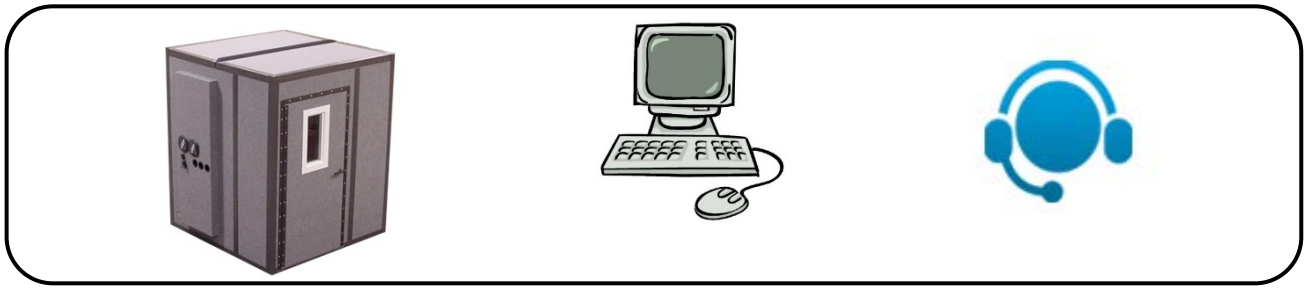
1. Background: Intelligibility evaluation

- Speech signal can suffer many types of degradation (e.g., noise, reverberant, channel distortion in speech communication), which may affect its intelligibility.
  - Intelligibility: Given a speech, how much could listeners understand? What is the accuracy rate?
- Speech intelligibility evaluation is an important topic.
  - To understand the factors accounting for speech intelligibility in various conditions
  - To design new speech proceesing (e.g., speech enhancement) methods
- Speech quality evaluation is not the target of this tutorial.

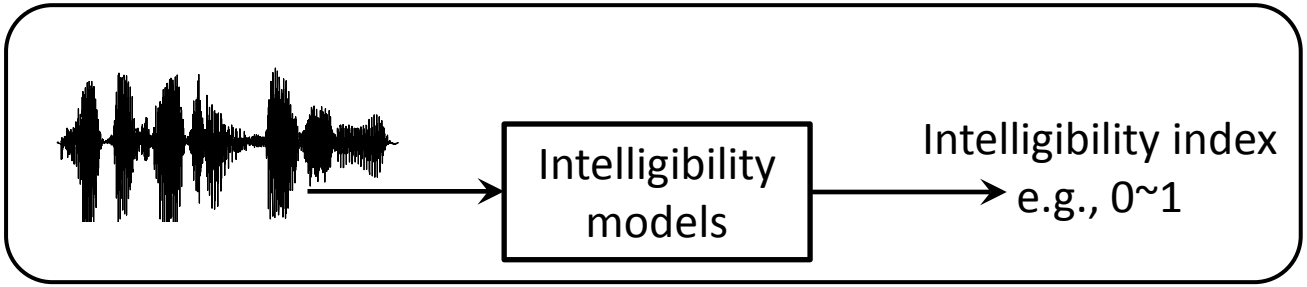


Intelligibility evaluation

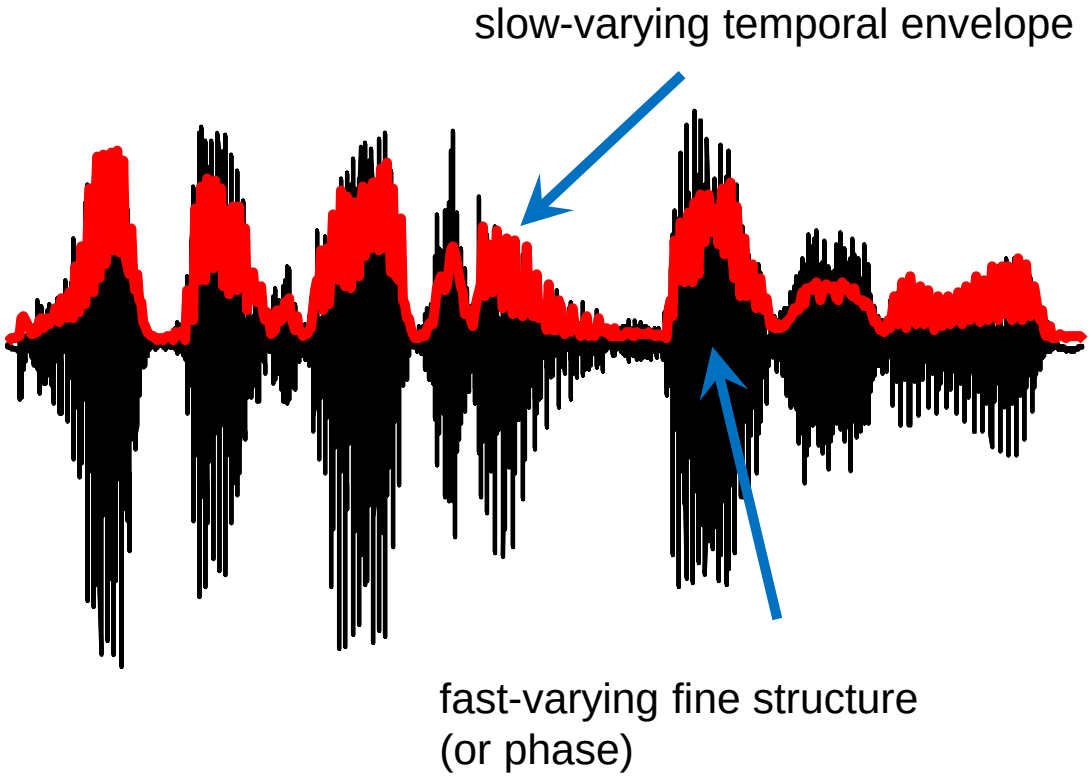
- The prediction of speech intelligibility can be divided into two categories: subjective prediction and objective prediction.
- Subjective speech intelligibility evaluation is the subjective reflection of the speech intelligibility.
  - most accurate
  - expensive, laborious, time-intensive, unsuitable for real-time monitoring purposes, and need specially-designed test materials.



- Objective speech intelligibility is developed based on models simulating human speech perception process.

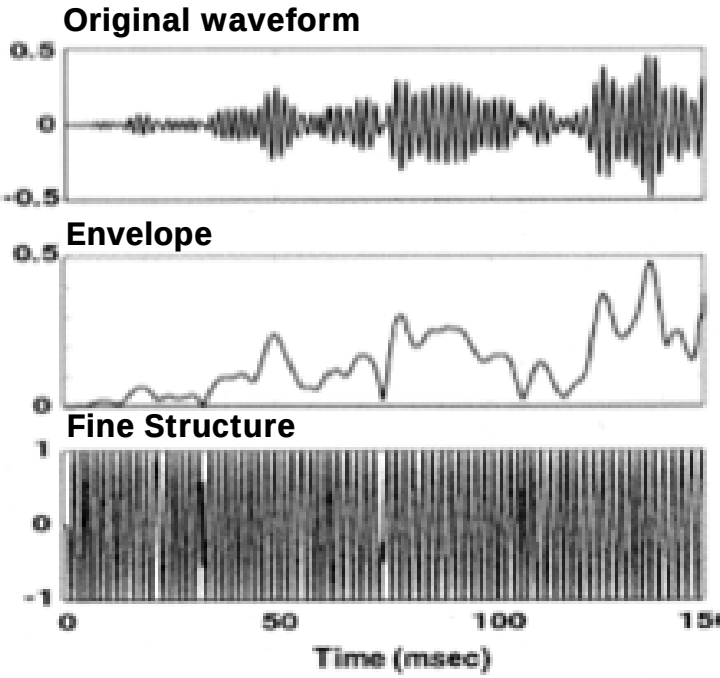


Background: Acoustic cues for speech perception



Hilbert transform

Signal= envelope × fine structure



Wilson et al., Ear Hear, 2005

Speech perception with temporal envelope

Speech Recognition with Primarily Temporal Cues

Robert V. Shannon,* Fan-Gang Zeng, Vivek Kamath, John Wygonski, Michael Ekelid

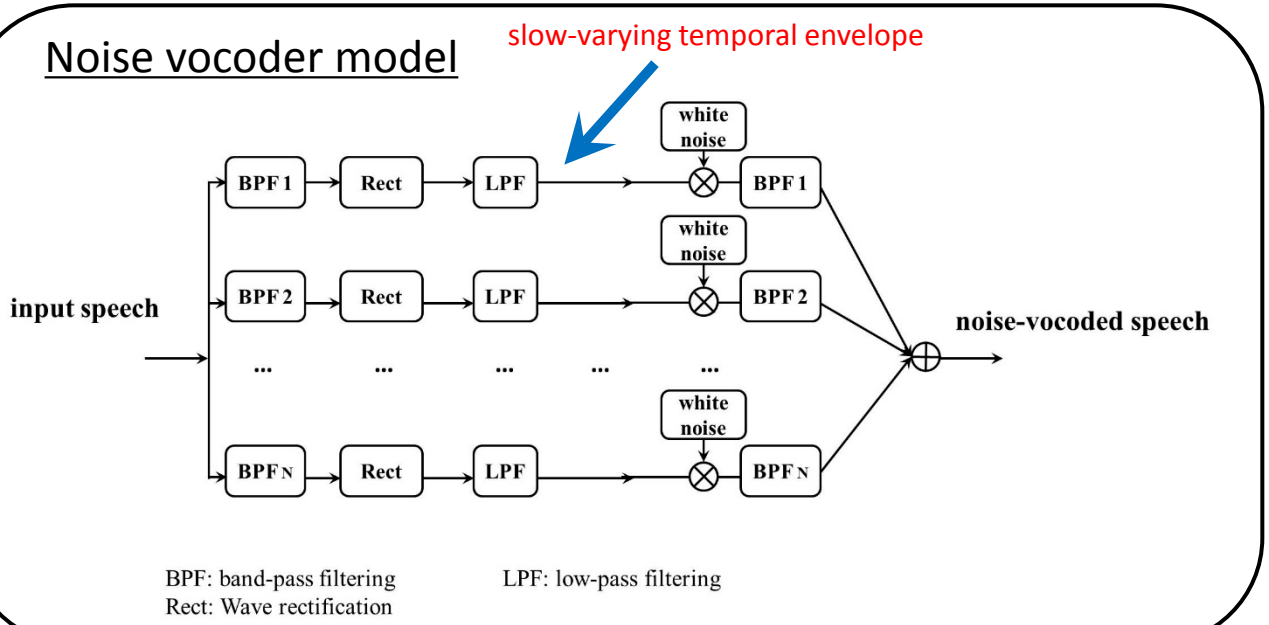
Nearly perfect speech recognition was observed under conditions of greatly reduced spectral information. Temporal envelopes of speech were extracted from broad frequency bands and were used to modulate noises of the same bandwidths. This manipulation preserved temporal envelope cues in each band but restricted the listener to severely degraded information on the distribution of spectral energy. The identification of consonants, vowels, and words in simple sentences improved markedly as the number of bands increased; high speech recognition performance was obtained with only three bands of modulated noise. Thus, the presentation of a dynamic temporal pattern in only a few broad spectral regions is sufficient for the recognition of speech.

under conditions of reduced spectral cues, slowly varying temporal information (<50 Hz) can yield relatively high speech recognition performance. This result is consistent with the observation of poor speech discrimination in children who have central processing disorders that disrupt temporal processing in the 20- to 50-ms range (*10*).

The specific reception of three speech features—voicing, manner, and place of articulation—was evaluated by information transmission analysis (*11*) on the consonant confusion matrix (Fig. 3). Information received on voicing and manner increased from one to two bands, to >90%, with no further improvement as the number of bands increased to three or four. Thus, bi-

Shannon et al., Science, 1995

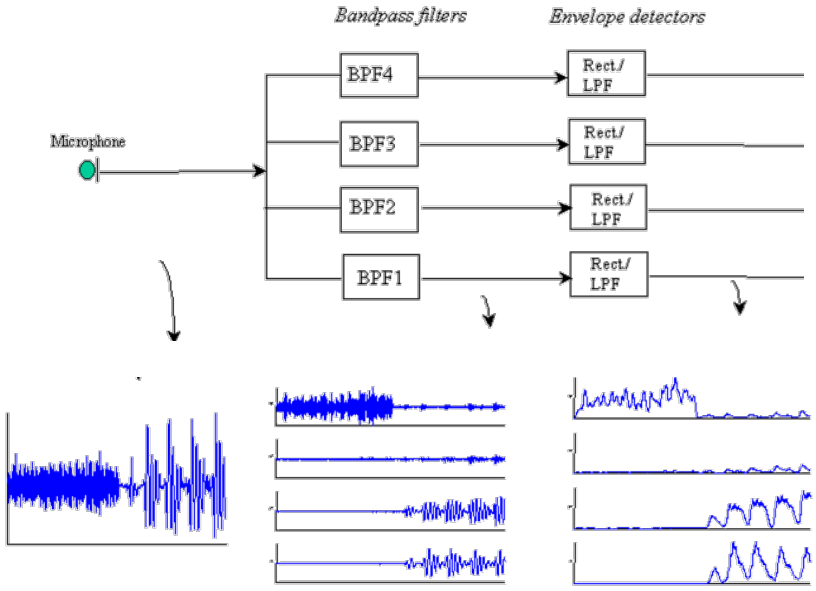
Noise vocoder model



Chen et al., JASA, 2017

Speech perception with temporal envelope

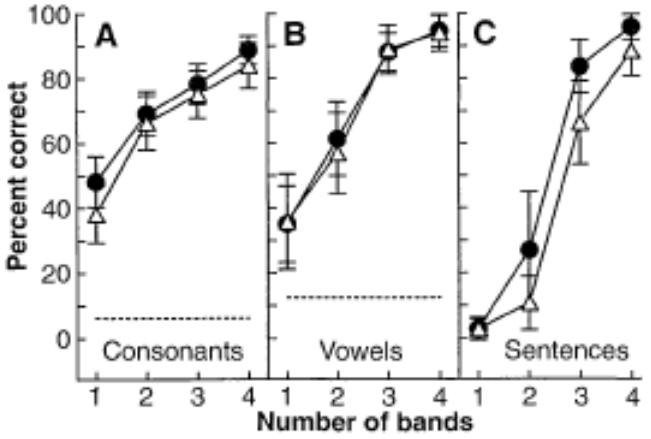
- In quiet, the envelope information from 4 bands was adequate to produce high levels of speech intelligibility to normal-hearing listeners.



Loizou, IEEE-EMB, 1999

# of channels	Wideband	Noise vocoded
N=1		
4		
8		

"The wife helped her husband"



Shannon et al., Science, 1995

More … (1)

1. Smith et al., Chimaeric sounds reveal dichotomies in auditory perception, Nature, 2002.
2. Zeng et al., On the dichotomy in auditory perception between temporal envelope and fine structure cues, JASA, 2004.
3. Xu et al., Relative contributions of spectral and temporal cues for phoneme recognition, JASA, 2005.
4. Gilbert et al., The ability of listeners to use recovered envelope cues from speech fine structure, JASA, 2006.
5. Lorenzi et al., Speech perception problems of the hearing impaired reflect inability to use temporal fine structure, PNAS, 2006.
6. Kazama et al., On the significance of phase in the short term Fourier spectrum for speech intelligibility, JASA, 2010.
7. Gonzalez et al., Gender and speaker identification as a function of the number of channels in spectrally reduced speech, JASA, 2005.
8. Chen et al., Effects of noise suppression and envelope dynamic range compression to the intelligibility of vocoded sentences for a tonal language, JASA, 2017.
9. Hazrati et al., The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners, Int J Audiology, 2012.

Background: Factors affecting human speech intelligibility

- Many important factors affect speech intelligibility, and accordingly speech intelligibility evaluation as well;
- So far, our knowledge to these effects is still limited:
  - Environment: noise, reverberation, and their combination.
  - Signal processing: e.g., dynamic range compression in hearing aids
  - Acoustic cues: temporal vs. spectral
  - Language: tonal vs. non-tonal
  - Listener: normal hearing vs. hearing impaired

Table of contents

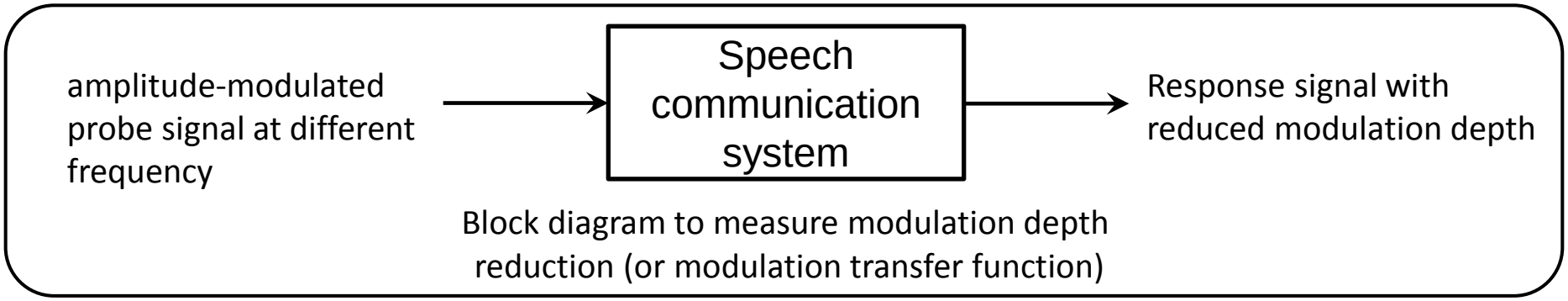
History: Intelligibility prediction models

- Historically, two lines of research serve as the foundation for existing intelligibility prediction models:
  - i) the Articulation Index (AI) by French and Steinberg, which was later refined and standardized as the Speech Intelligibility Index (SII),
  - ii) the Speech Transmission Index (STI) by Steeneken and Houtgast.
- AI and SII were developed with simple linear signal degradations (e.g., additive noise).
  - divide the signal under analysis into frequency subbands, and assume that each subband contributes independently to intelligibility.
  - estimating the long-term speech and noise power within the subband to arrive at the long-term subband signal-to-noise ratio (SNR).
  - The subband SNRs are limited to the range [−15, +15 dB], normalized to a value between 0 and 1, and combined as a perceptually weighted average.
  - Limitations:
    - require the long-term spectrum of the additive noise signal to be known in advance.
    - cannot discern modulated noise signals from un-modulated ones, when their long-term spectra are identical.
    - AI and SII are not directly applicable to signals which have been passed through some non-linear processing stage before presented to the listener.

History: Intelligibility prediction models

- Extended SII (ESII, Rhebergen and Versfeld, 2005) avoids the use of long-term noise spectra in SII, and divides the masker signal into short time frames (9–20 ms) and averages the SII computed for each frame, to predict intelligibility for fluctuating noise sources.
- Coherence SII (CSII, Kates and Arehart, 2005) extends SII to better take into account various non-linear distortions, including center- and peak-clipping.

- STI extends the range of distortions to convolutive noise (e.g., reverberant speech and effects of room acoustics).
  - based on the observation that reverberation and/or additive noise tend to reduce the depth of temporal signal modulations compared to the clean, undistorted reference signal.

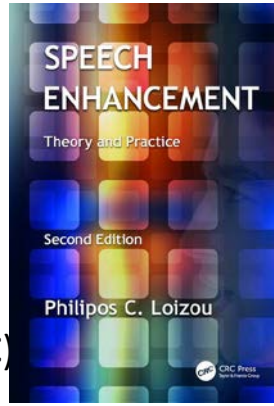


Block diagram to measure modulation depth reduction (or modulation transfer function)

- Speech-based STI (Goldsworthy and Greenberg, 2004) replaces the traditional noise probe signals with actual speech signals to better take into account the effect of non-linear distortions (e.g., envelope clipping, dynamic amplitude compression, etc.)

Intelligibility data 1: noise-suppressed speech

- IEEE sentences (in English), and normal-hearing listeners
- 4 types of masker: babble, car, street, train
- 8 commonly-used single-channel noise-suppression algorithms:
  1. the generalized subspace approach (i.e., KLT)
  2. the perceptually-based subspace approach (i.e., KLT_Jab)
  3. the log minimum mean square error algorithm (i.e., logMMSE)
  4. the logMMSE algorithm with speech-presence uncertainty (i.e., logMMSE_spu)
  5. the spectral subtraction algorithm based on reduced-delay convolution (i.e., RDC)
  6. the multiband spectral-subtractive algorithm (i.e., MB)
  7. the Wiener filtering algorithm based on wavelet-thresholded multitaper spectra (i.e., Wiener_thr)
  8. the traditional Wiener algorithm (i.e., Wiener)
- A total of 72 conditions
  - 4 maskers × 2 SNR levels × 8 algorithms + 4 maskers × 2 noisy references (i.e., 2 noise-masked conditions at 2 SNR levels).



Loizou, 2007; Ma et al., JASA, 2009; Chen et al., Speech communication 2016; Van Kuyk et al., IEEE/ACM TASPL 2018

Intelligibility data 2: noise masking

- Subjects: 8 normal-hearing native-Mandarin speakers
- Materials: Mandarin version of Hearing in noise Test (MHINT)
- 22 noise-masking conditions:

SNR levels chosen for the various maskers used.

Masker	SNR (dB)
Babble	−10, −5, −3, 0, 5
Car	−15, −10, −5
SSN	−15, −10, −7, −5, −3, 0, 5
Street	−15, −10, −7, −5, −3, 0, 5

Intelligibility data 3: reverberant speech

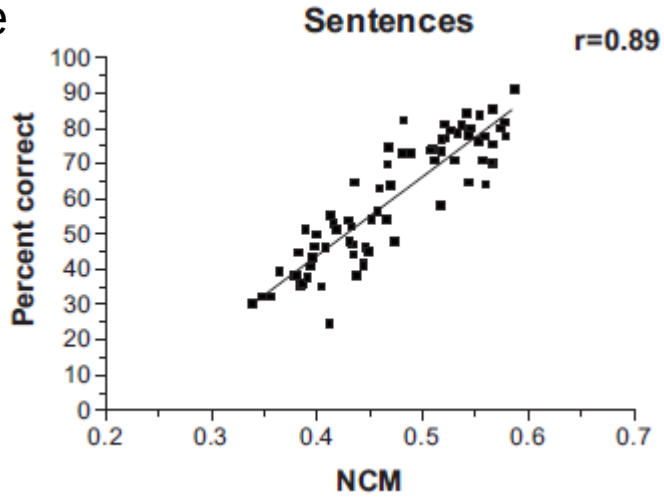
- Subjects: 11 cochlear-implant users
- Materials: IEEE database
- A total of 21 test conditions:
  - 1 anechoic (quiet) condition
  - 4 reverberant (T_{60} = 0.3, 0.6, 0.8, and 1.0 s) conditions
  - 4 reverberant + noisy (combinations of T_{60} = 0.6 and 0.8 s with SNR = 5 and 10 dB) conditions
  - 12 conditions involving reverberant sentences processed via an ideal reverberant mask algorithm (in T_{60} = 0.6 and 0.8 s and SNR levels of 5 and 10 dB using three different binary mask threshold values of −8, −10 and −12 dB)

Hazrati et al., IJA, 2012; Falk et al., IEEE Sig. Proc. Mag., 2015; Chen et al., BSPC, 2013

Performance evaluation

- Metrics:(between the predictions and real scores)
  - (Pearson or Kendall's Tau) correlation coefficient (*r*)
  - Root Mean Square Error (RMSE)

Example



$$y = \frac{1}{1+e^{-(b_1 \times index - b_2)}} \times 100$$

(b_1, b_2) are the fitting parameters
Compute the correlation between predicted scores and real scores

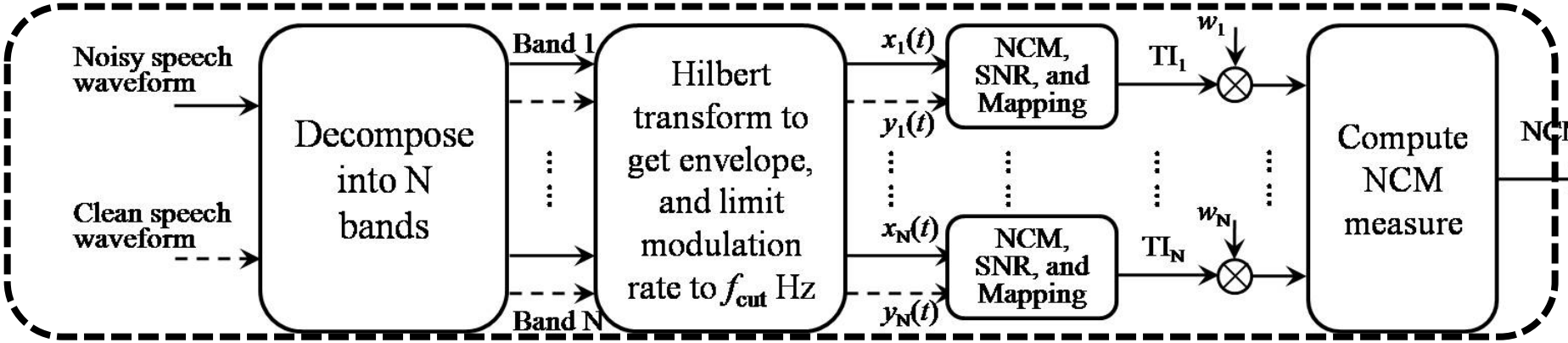
Example

TABLE III
PERFORMANCE METRICS FOR THE FIVE SIP ALGORITHMS

SIP algorithm	RMSE	Kendall's Tau
CNN	17.69 pp	0.667
STOI	18.94 pp	0.658
ESTOI	17.11 pp	0.692
NI-STOI	19.90 pp	0.629
SRMR	32.77 pp	0.281

2.2.1. Speech-based STI: Normalized-covariance measure (NCM)

- Envelope-based intelligibility index



$$NCM = \frac{\sum_{i=1}^{N} TI_i \times w_i}{\sum_{i=1}^{N} w_i}$$

$$r_i = \frac{\Sigma_t (x_i(t) - \mu_i)(y_i(t) - \nu_i)}{\sqrt{\Sigma_t (x_i(t) - \mu_i)^2}\sqrt{\Sigma_t (y_i(t) - \nu_i)^2}}, \quad |r_i| \leq 1$$

$$SNR_i = 10 \log_{10}\left(\frac{r_i^2}{1 - r_i^2}\right)$$

limited to the range of $[-15, 15]$ dB

$$TI_i = \frac{SNR_i + 15}{30}$$

(TI: transmission index, and ≤ 1)

Goldsworthy et al., JASA, 2004

Envelope modulation frequency (f_{cut}) and band weighting (w_i)

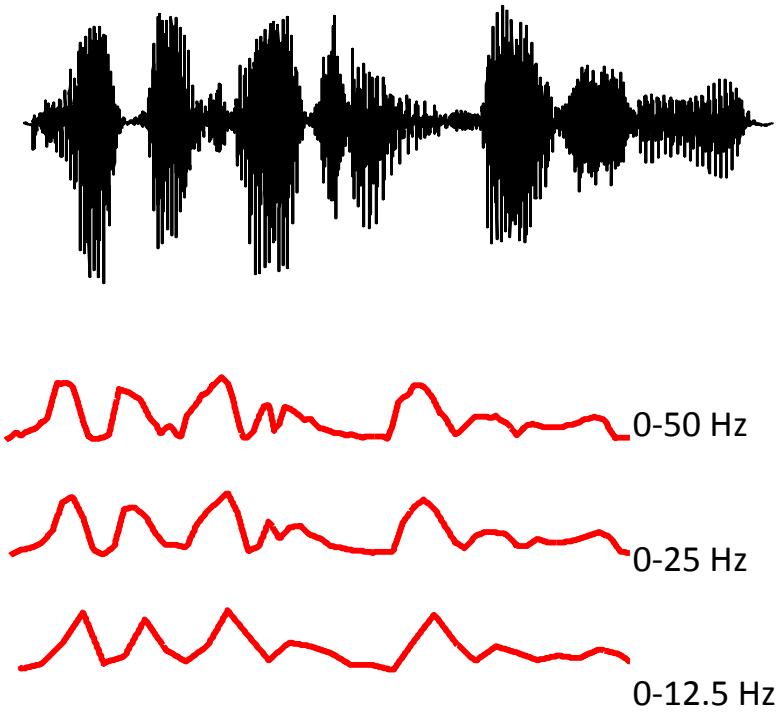
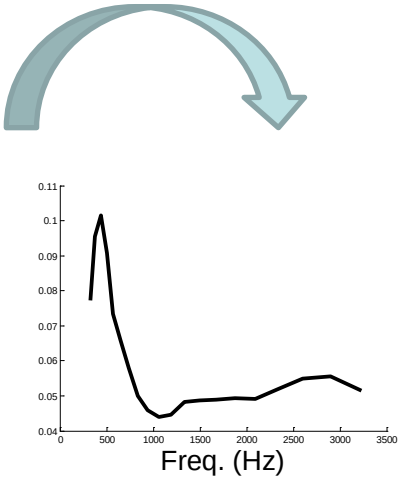


TABLE II. AI weights (ANSI, 1997) used in the implementation of the NCM measure for consonants and sentence materials.

Band	Center freq. (kHz)	Consonants	Sentences
1	0.3249	0.0346	0.0772
2	0.3775	0.0392	0.0955
3	0.4356	0.0406	0.1016
4	0.5000	0.0420	0.0908
5	0.5713	0.0433	0.0734
6	0.6502	0.0457	0.0659
7	0.7376	0.0472	0.0580
8	0.8344	0.0473	0.0500
9	0.9416	0.0471	0.0460
10	1.0602	0.0487	0.0440
11	1.1915	0.0519	0.0445
12	1.3370	0.0534	0.0482
13	1.4980	0.0562	0.0488
14	1.6763	0.0612	0.0488
15	1.8737	0.0684	0.0493
16	2.0922	0.0732	0.0491
17	2.3342	0.0748	0.0520
18	2.6022	0.0733	0.0549
19	2.8989	0.0685	0.0555
20	3.2274	0.0670	0.0514



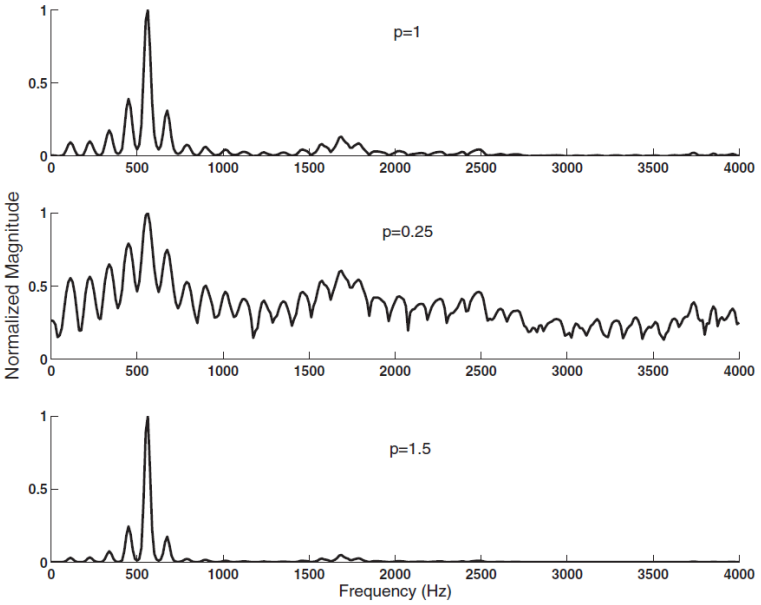
American National Standards Institute (ANSI), 1997

Signal-dependent band weighting (w_i)

$$W(j,m) = X(j,m)^p$$

where *p* is the power exponent, and can be varied for maximum correlation and optimized for different materials.

- Values of $p<1$ compress the spectrum, while values of $p>1$ expand the spectrum.
- The value of *p* can control the emphasis or weight placed on spectral peaks and/or spectral valleys.
- Compressive values of *p* ($p<1$) equalize the spectrum by boosting the low-intensity components, e.g., spectral valleys.



Ma et al., JASA, 2009

Signal-dependent band weighting (w_i)

$$W_i^{(1)} = \left(\sum_t x_i^2(t) \right)^p, \qquad (12)$$

$$W_i^{(2)} = \left(\sum_t (\max[x_i(t) - d_i(t), 0])^2 \right)^p, \qquad (13)$$

where $d_i(t)$ denotes the downsampled masker signal in the time domain. The power exponent p was varied from 0.12 to 1.5.

- The motivation behind $W_i^{(1)}$ is to place weight to each TI value in proportion to the signal energy in each band.
- The motivation behind $W_i^{(2)}$ is to place weight to each TI value in proportion to the excess masked signal.

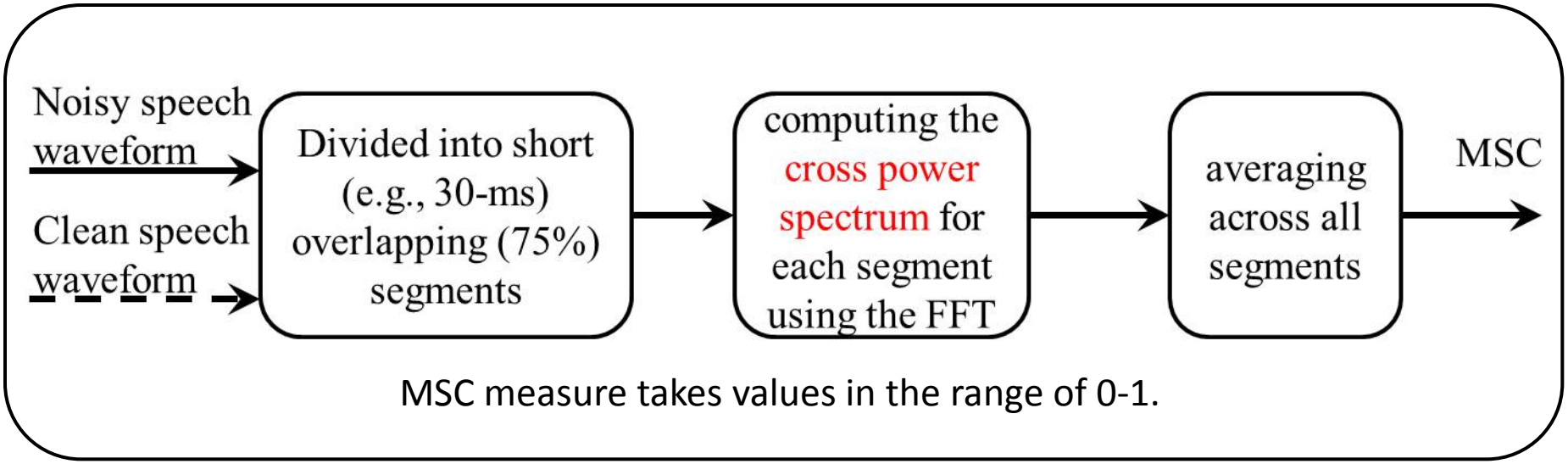
Intelligibility data 1: noise-suppressed speech

		r
NCM	ANSI (Table II)	0.82
NCM	$W_i^{(1)}$, $p=1.5$, Eq. (12)	0.89
NCM	$W_i^{(2)}$, $p=1.5$, Eq. (13)	0.89

Ma et al., JASA, 2009

2.2.2. The coherence-based speech intelligibility index (CSII)

- Spectral-detail based index
- Compute the magnitude-squared coherence (MSC) function



➢ For M data segments (frames), the MSC at frequency bin ω is given by

$$MSC(\omega) = \frac{\left|\sum_{m=1}^{M} X_m(\omega) Y_m^*(\omega)\right|^2}{\sum_{m=1}^{M} |X_m(\omega)|^2 \sum_{m=1}^{M} |Y_m(\omega)|^2}$$

$X_m(\omega)$ and $Y_m(\omega)$ denote the FFT spectra of the $x(t)$ and $y(t)$ signals, computed in the mth segment.

The coherence-based speech intelligibility index (CSII)

- Define a signal-to-noise ratio

$$\mathrm{SNR}_{CSII}(j,m) = 10\log_{10}\frac{\sum_{k=1}^{N} G_j(\omega_k)\times MSC(\omega_k)\left|Y_m(\omega_k)\right|^2}{\sum_{k=1}^{N} G_j(\omega_k)\times[1-MSC(\omega_k)]\left|Y_m(\omega_k)\right|^2} \quad (1)$$

$G_j(\omega)$ denotes the ro-ex filter centered around the *j*th critical band

$Y(\omega_k)$ is the FFT spectrum of the processed signal, and *N* is the FFT size

- SNR term was limited to [−15, 15] dB, and mapped linearly between 0 and 1, as:

$$T_{CSII} = (SNR_i + 15)/30. \quad (2)$$

- The coherence-based speech intelligibility index (CSII):

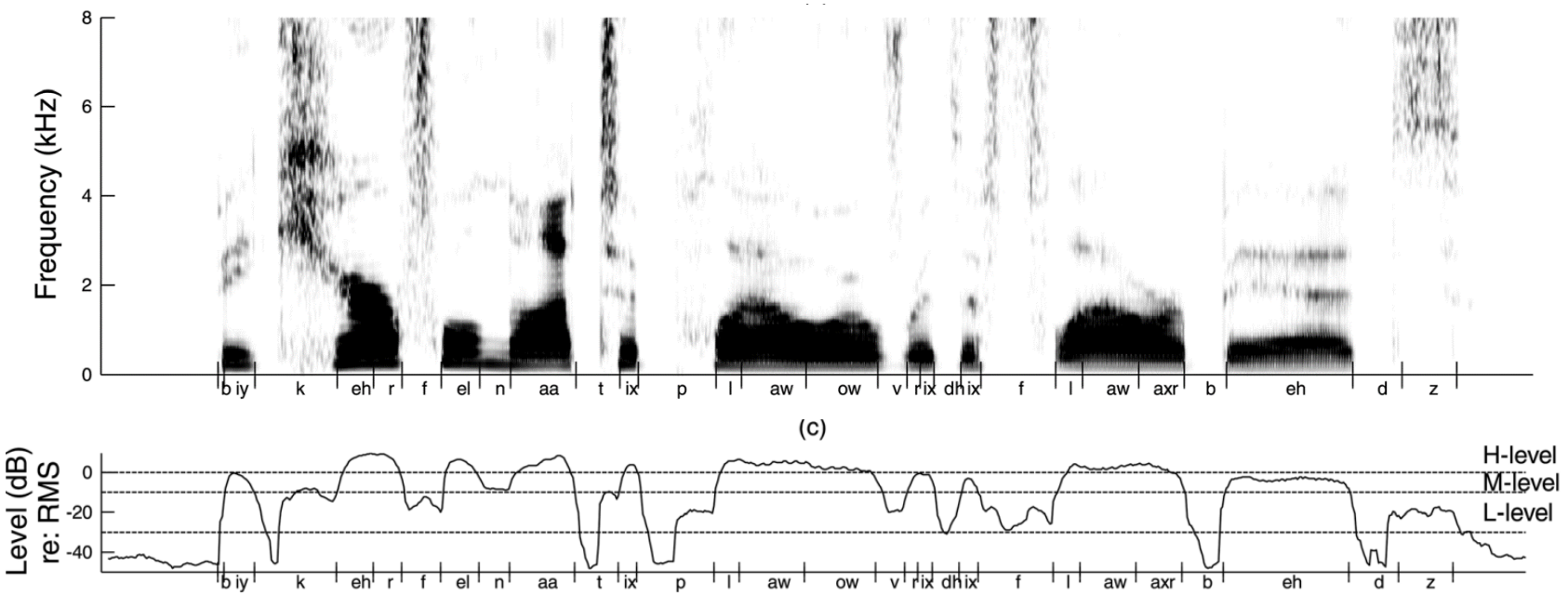
$$\mathrm{CSII} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W(j)T_{CSII}(j,m)}{\sum_{j=1}^{K} W(j)}, \quad (3)$$

where *K* is the number of frequency bands, and

$W(j)$ is the weight placed on the *j*th frequency band.

Kates, JASA, 1992; Kates et al., 2005

CSII: Three level (high, middle, low) contribution



0dB: overall level of the sentence

- H-level: > overall RMS level
- M-level: [overall RMS-10 dB, over all RMS]
- L-level: [overall RMS-30 dB, overall RMS-10 dB]

Kates et al., JASA, 2005

CSII: Three level (high, middle, low) contribution

- For most part,
  - H-level segments: vowels and semi-vowels
  - M-level segments: consonants and vowel-consonant transitions
  - L-level segments: primarily weak consonants

IEEE sentence

	Vowels	Consonants	Non-phoneme
H-level	79.5%	20.5%	0.0%
M-level	46.2%	53.7%	0.1%
L-level	6.8%	84.5%	8.7%

	RMS-level		
	H-level	M-level	L-level
C-V boundaries	18.9%	53.6%	27.5%

Intelligibility data 1: noise-suppressed speech

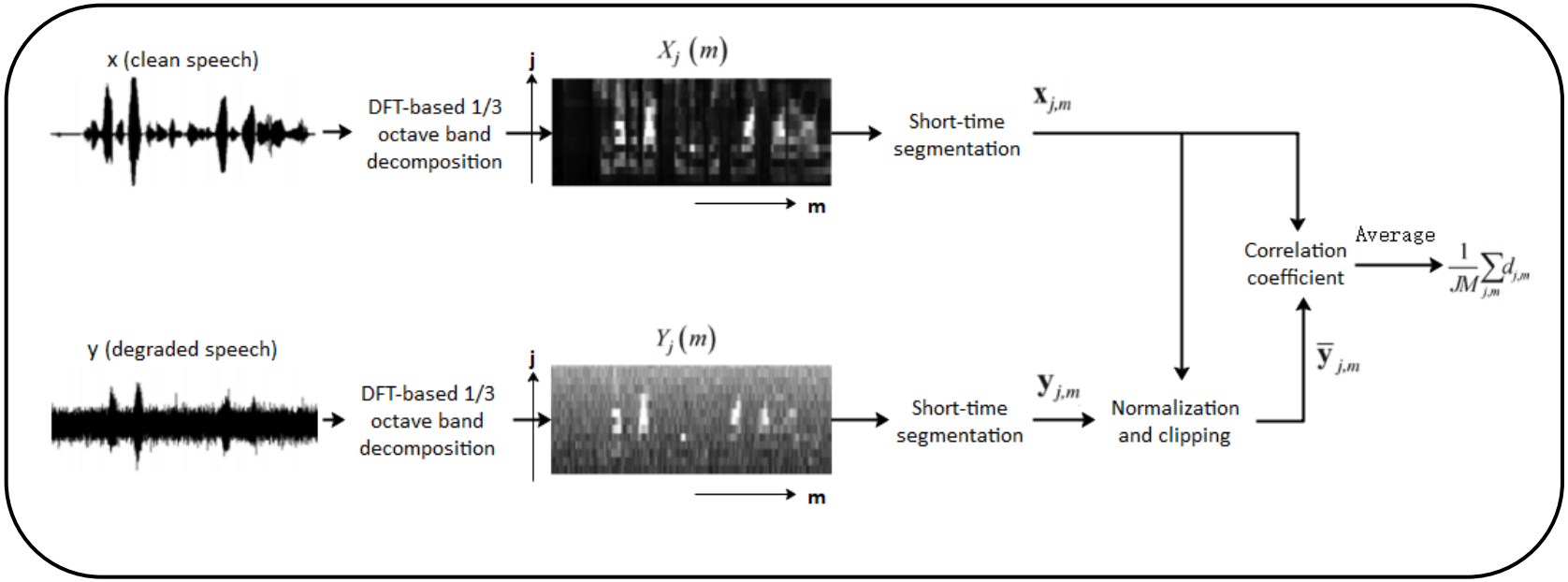
		RMS-level		
Intelligibility measure	All segments	H-level	M-level	L-level
NCM	0.80	0.83	0.89	0.77
CSII	0.82	0.85	0.91	0.86

Kates et al., JASA, 2005; Chen et al., JASA, 2012

2.2.3. Short-time objective intelligibility (STOI)

STOI is a function of a time-frequency dependent intermediate intelligibility measure

- Compare the temporal envelopes of the clean and degraded speech in short-time regions by means of a correlation coefficient.



Taal et al., IEEE/ACM TASLP, 2011

1) decomposing clean (x) and degraded (y) signals into the time-frequency units $[X_j(m), Y_j(m)]$ by the discrete Fourier transform-based one-third octave band

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} |\hat{x}(k,m)|^2} \quad (1)$$

2) calculating the short-time temporal envelope of the time-frequency representation $[x_{j,m}, y_{j,m}]$ of clean and degraded signals

$$x_{j,m} = \left[X_j(m-N+1), X_j(m-N+2), \ldots, X_j(m)\right]^T \quad (2)$$

3) normalizing and clipping the short-time temporal envelope of the degraded speech, represented as $\bar{y}_{j,m}$

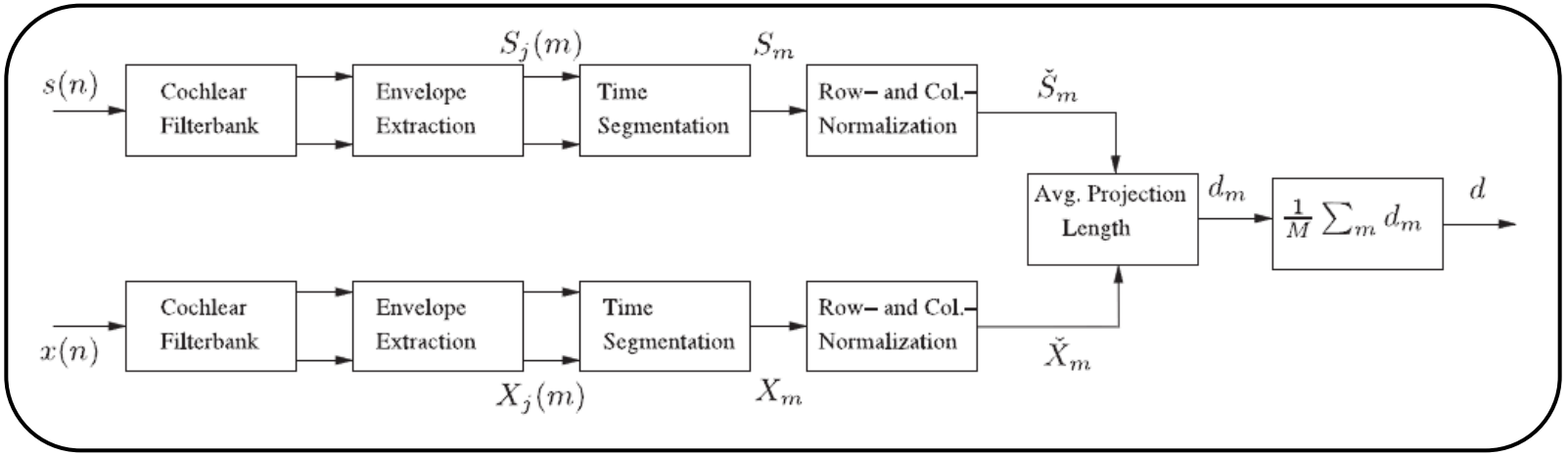
$$\bar{y}_{j,m}(n) = \min\left(\frac{\|X_{j,m}\|}{\|Y_{j,m}\|} y_{j,m}(n), \left(1 + 10^{-\frac{\beta}{20}}\right) x_{j,m}(n)\right) \quad (3)$$

4) comparing the temporal envelopes of the clean and degraded speech in short-time regions by means of a correlation coefficient

$$d_{j,m} = \frac{(x_{j,m} - \mu_{x_{j,m}})^T (\bar{y}_{j,m} - \mu_{\bar{y}_{j,m}})}{\|x_{j,m} - \mu_{x_{j,m}}\| \|\bar{y}_{j,m} - \mu_{\bar{y}_{j,m}}\|}, \mathrm{d} = \frac{1}{JM} \sum_{j,m} d_{j,m} \quad (4)$$

extended STOI (ESTOI)

- STOI assumes mutual independence between frequency bands, and independent frequency band contributions to intelligibility.
- ESTOI shares the first step with STOI, i.e., mean- and variance-normalization is applied to subband envelope.
- ESTOI also computes the *spectral* correlation across frequency bands, which are finally averaged across time.
  - Better capture the effect of time-modulated noise maskers.



Block diagram of ESTOI measure

Jensen and Taal, IEEE/ACM TASLP, 2016

Short-time spectrogram matrix

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

$$s_{j,m} = [S_j(m-N+1)\, S_j(m-N+2) \cdots S_j(m)]^T$$

STOI:
the jth mean- and variance normalized row of S_m

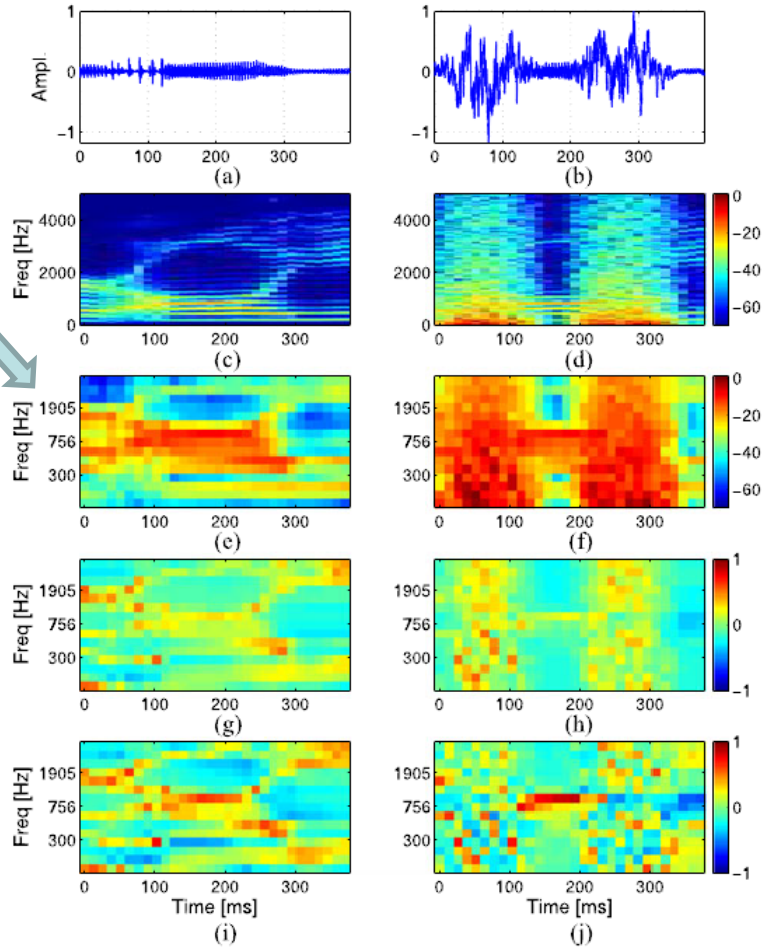
$$\bar{s}_{j,m} = \frac{1}{\|(s_{j,m} - \mu_{s_{j,m}} \underline{1})\|} (s_{j,m} - \mu_{s_{j,m}} \underline{1})$$

ESTOI:
the row-normalized spectrogram matrix

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

Normalize the nth column, and define the row- and column-normalized matrix

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



Jensen and Taal, IEEE/ACM TASLP, 2016

2.2.4. Hearing-Aid Speech Perception Index (HASPI)

- Based on a model of the auditory periphery that incorporates changes due to hearing loss.
- Compare the envelope and temporal fine structure outputs of the auditory model for a reference signal to the outputs of the model for the signal under test.
  - The auditory model for the reference signal is set for normal hearing, while the model for the test signal incorporates the peripheral hearing loss.

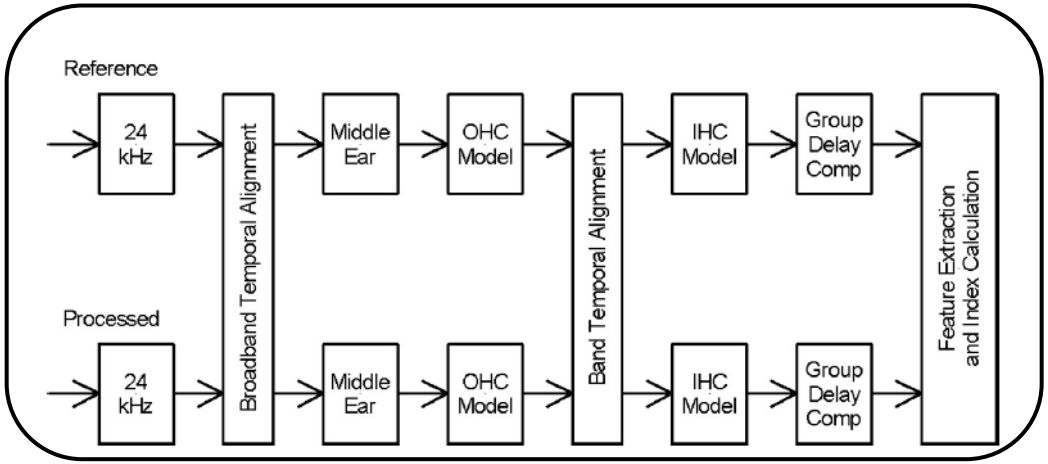


Fig. 1. Block diagram showing the reference and processed signal comparison.

OHC: outer hair cell
IHC: inner hair cell

Kates and Arehart, Speech Communication, 2014

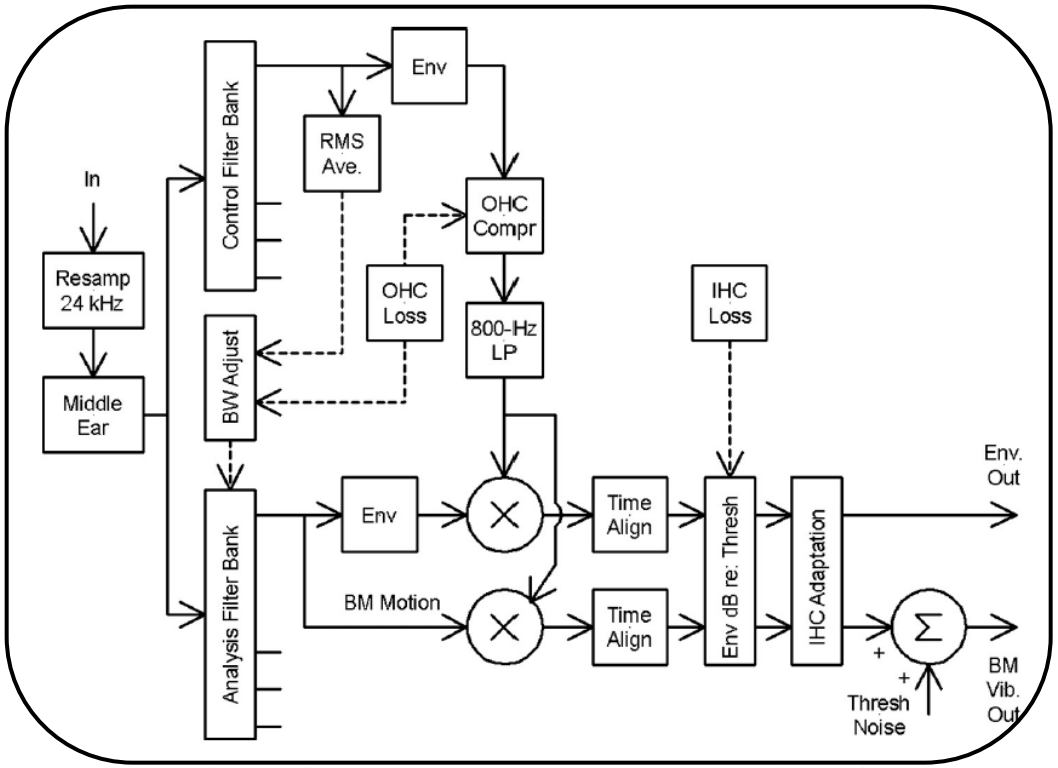


Fig. 2. Block diagram of the auditory model used to extract the signals in each frequency band.

*C*epstral correlation with envelope

*A*uditory coherence with temporal fine-structure

$$p = -9.047 + 14.817c + 0.0a_{Low} + 0.0a_{Mid} + 4.616a_{High}$$

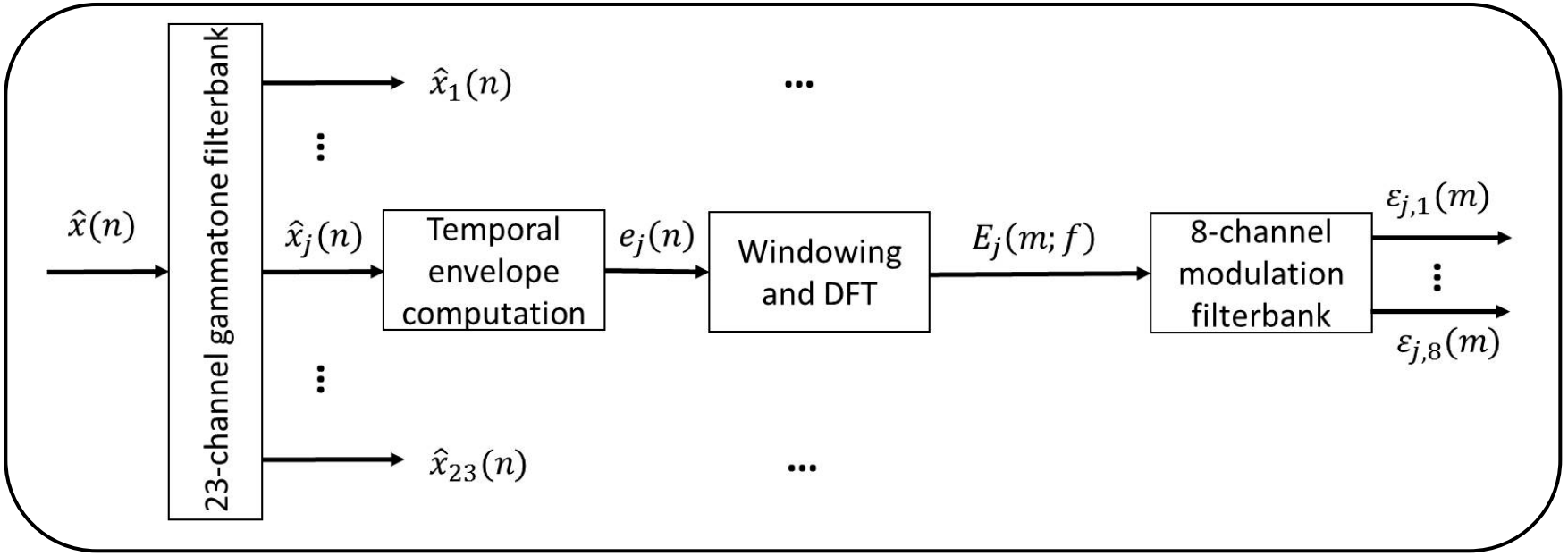
HASPI is found to give accurate intelligibility predictions for a wide range of signal degradations, including:

- speech degraded by noise and nonlinear distortion,
- speech processed using frequency compression,
- noisy speech processed through a noise-suppression algorithm, and
- speech where the high frequencies are replaced by the output of a noise vocoder.

Kates and Arehart, Speech Communication, 2014

2.2.5. Speech-to-reverberation modulation energy ratio (SRMR)

- Most of the previous indices are intrusive. That is, a clean reference is needed.
  - In most scenarios, such reference signal is not available.
  - Non-intrusive (or reference-free) index only uses degraded speech signal to predict intelligibility.
- The SRMR measure is computed by performing spectral analysis on the modulation envelopes of the response (e.g., reverberant) speech signal.



Tiago et al., IEEE/ACM TASLP, 2010

modulation envelopes

1) The (de)reverberant speech signal $\hat{x}(n)$ is filtered by a 23-channel gammatone filterbank to emulate the processing performed by the cochlea.
2) The temporal envelope $e_j(n)$of the j^{th} filter output signal $\hat{x}_j(n)$ is computed using the Hilbert transform $\mathcal{H}\{\cdot\}$, where *j*=1, …, 23.
3) Modulation spectral energy for critical band j is then computed as the squared magnitude of the discrete Fourier transform $\mathcal{F}\{\cdot\}$ of the temporal envelope $e_j(m;n)$.

$$e_j(n) = \sqrt{\hat{x}_j(n)^2 + \mathcal{H}\{\hat{x}_j(n)\}^2} \quad (1)$$

$$E_j(m;f) = \left|\mathcal{F}(e_j(m;n))\right|^2 \quad (2)$$

4) The averaged modulation energy over all frames of the j^{th} critical-band signal grouped by the k^{th} modulation filter was calculated, represented as $\bar{\varepsilon}_k$., where *k* = 1, …, 8.

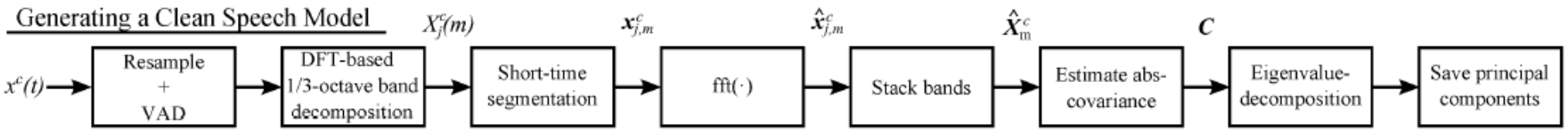
$$\bar{\varepsilon}_k = \frac{1}{23}\sum_{j=1}^{23}\bar{\varepsilon}_{j,k} \quad (3)$$

5) The SRMR is given by the speech to reverberation modulation energy ratio. The upper summation bound K^* in the denominator is adapted to the speech signal under test.

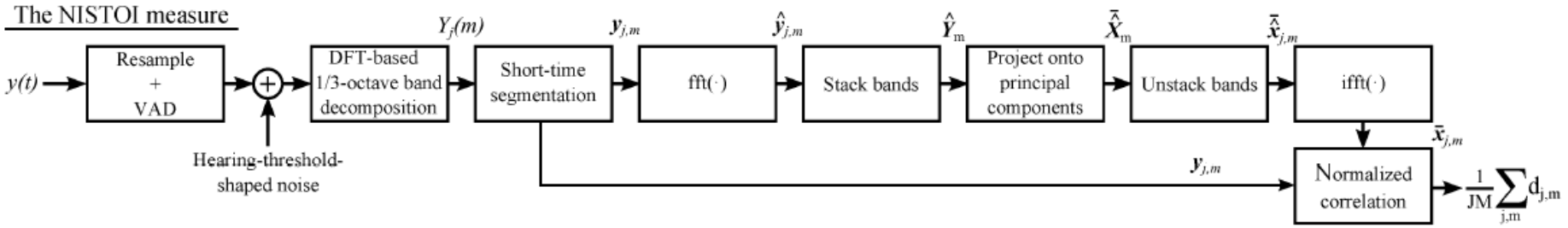
$$SRMR = \frac{\sum_{k=1}^{4}\bar{\varepsilon}_k}{\sum_{k=5}^{K^*}\bar{\varepsilon}_k} \quad (4)$$

Nonintrusive STOI measure (NISTOI)

- Using a statistical model of clean speech to estimate clean signal amplitude envelopes from the degraded signal.



- Subsequently, the STOI measure is evaluated by using the envelopes of the degraded signal and the estimated clean envelopes.

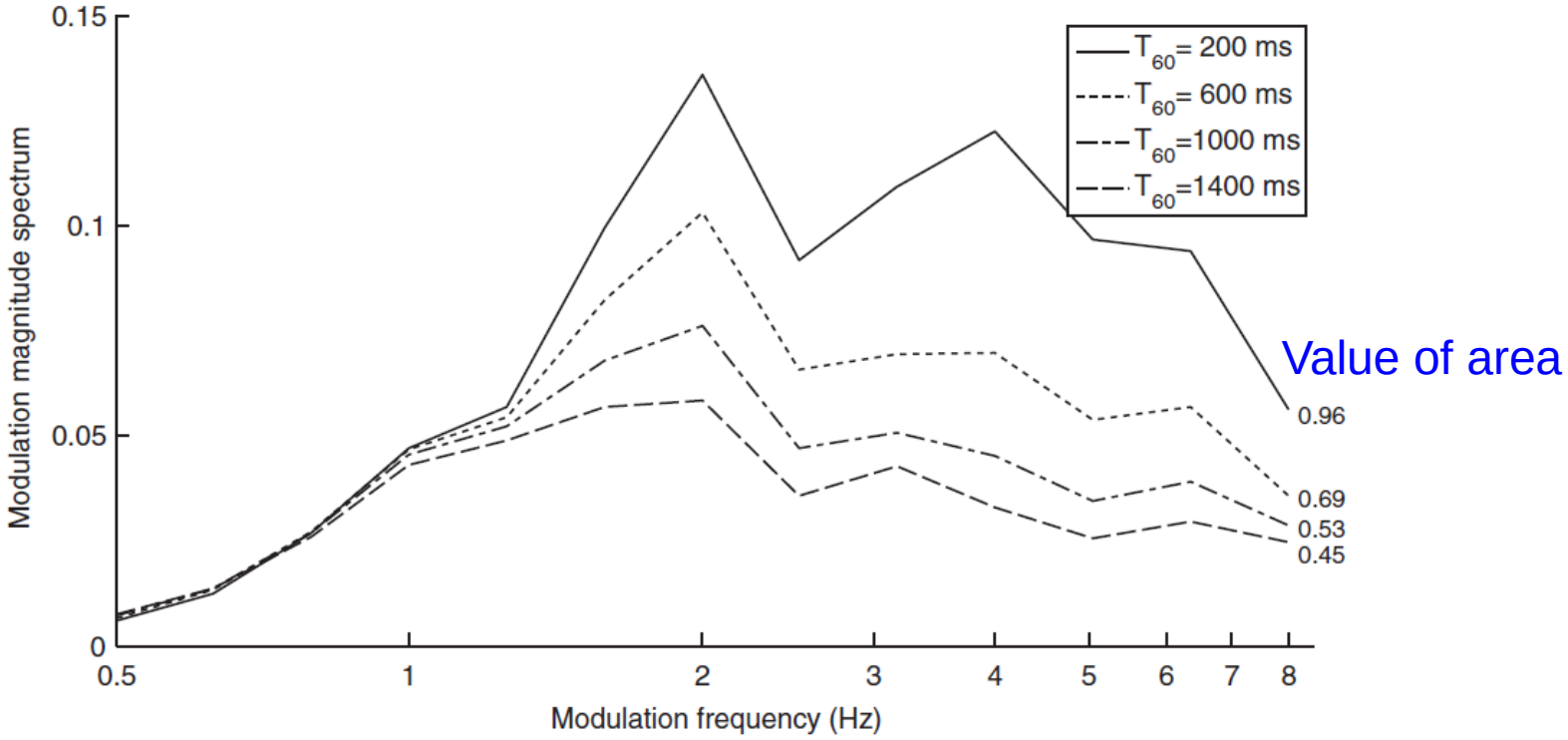


	r	RMSE
STOI [16]	0.959	9.4%
NI-STOI (Dantale II)	0.711	25.2%
NI-STOI (TIMIT F)	0.704	25.4%
NI-STOI (TIMIT M+F)	0.702	25.5%
SRMR [20]	0.237	45.2%
SRMR-norm [31]	0.394	38.6%

Andersen et al., ICASSP, 2017

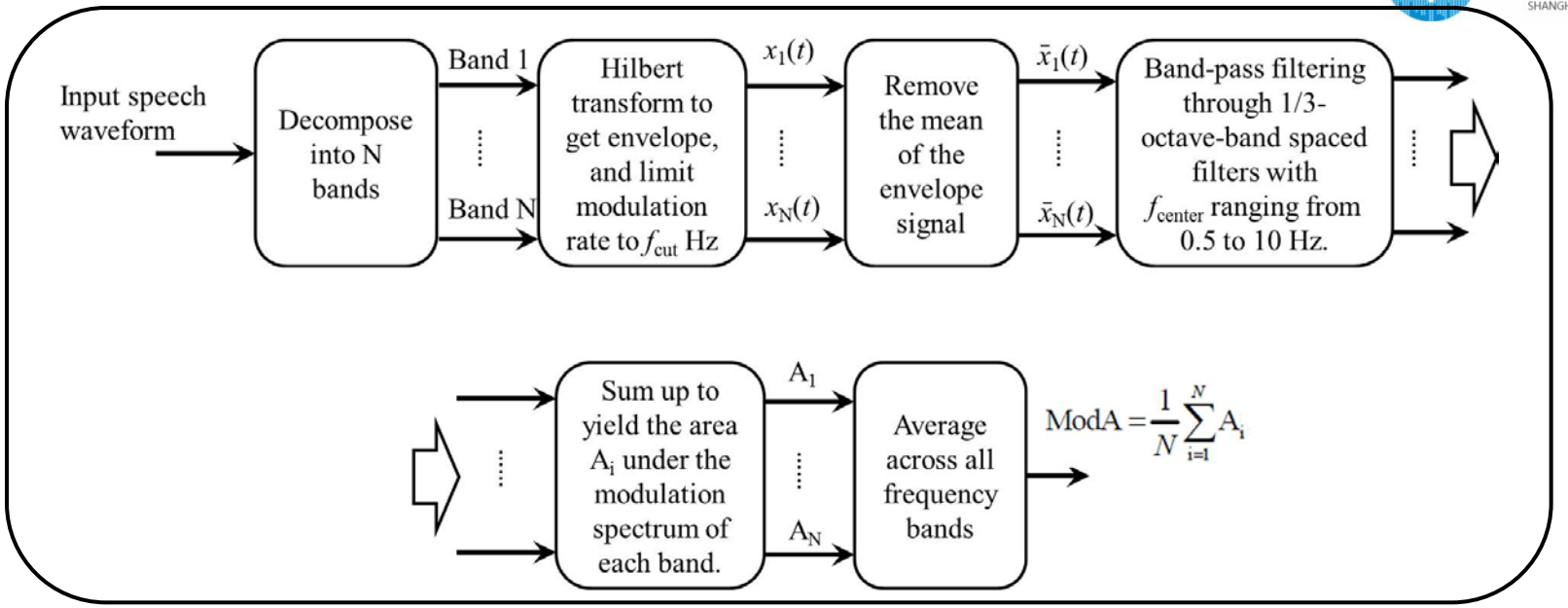
Nonintrusive: the Average Modulation-spectrum Area (ModA)

- As the level of reverberation is increased, the modulation spectrum of the reverberant envelopes becomes flat and shifts down.



Speech modulation spectra computed in four reverberant conditions for a frequency band spanning 775–1735 Hz。

Chen et al., BSPC, 2013



Intelligibility data 3: reverberant speech

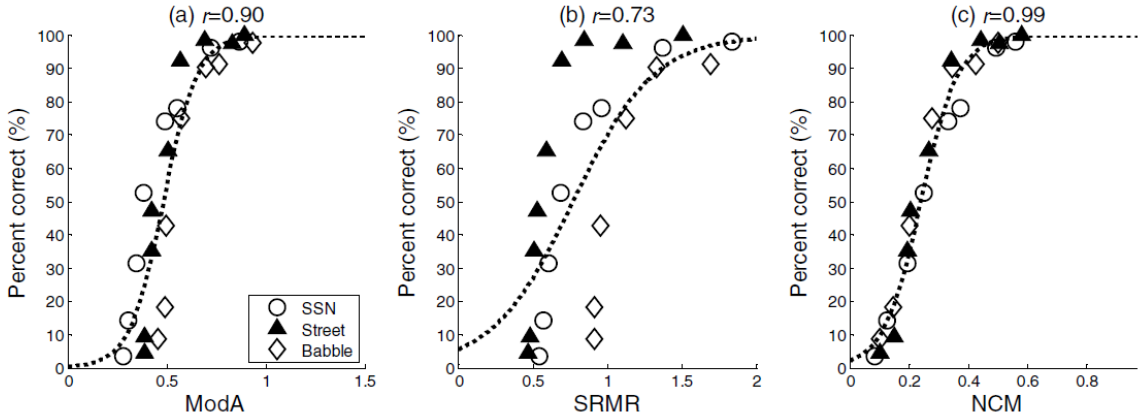


Fig. 2. Scatter plots of sentence recognition scores against the (a) ModA, (b) SRMR, and (c) NCM values.

Chen et al., BSPC, 2013

More … (2)

More index (intrusive):

1. Elhilali et al., A spectro-temporal modulation Index (STMI) for assessment of speech intelligibility, Speech Communication, 2003
2. Cooke, A glimpsing model of speech perception in noise, JASA, 2007.
3. Ma et al., SNR Loss: A new objective measure for predicting speech intelligibility of noise-suppressed speech, Speech Communication, 2010.
4. Jørgensen et al., Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing, JASA, 2011.
5. Gomez et al., Improving objective intelligibility prediction by combining correlation and coherence based methods with a measure based on the negative distortion ratio, Speech Communication, 2012
6. Chen et al., Modifying the normalized covariance metric measure to account for non-linear distortions introduced by noise-reduction algorithms, JASA, 2013.
7. Chen et al., A Hilbert-fine-structure-derived physical metric for predicting the intelligibility of noise-distorted and noise-suppressed speech, Speech Communication, 2013.
8. Mamun et al., Prediction of speech intelligibility using a neurogram orthogonal polynomial measure (NOPM), IEEE/ACM TASLP, 2015.
9. Lightburn et al., A weighted STOI intelligibility metric based on mutual information, ICASSP, 2016.
10. Chen et al., Modeling speech intelligibility with recovered envelope from temporal fine structure stimulus, Speech Communication, 2016.

More index (nonintrusive):

1. Chen, Modeling noise influence to speech intelligibility non-intrusively by reduced speech dynamic range, InterSpeech, 2016.
2. Chen, Predicting the intelligibility of noise-corrupted speech non-intrusively by across-band envelope correlation, BSPC, 2016.
3. Karbasi et al., Twin HMM-based non-intrusive speech intelligibility prediction, ICASSP, 2016.
4. Sørensen et al., Non-intrusive codebook-based intelligibility prediction, Speech Communication, 2018.

Effects of weighting function:

1. Ma et al., Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions, JASA, 2009.
2. Chen et al., Analysis of a simplified normalized covariance measure based on binary weighting functions for predicting the intelligibility of noise-suppressed speech, JASA, 2010.
3. Chen et al., Frequency importance function of the speech intelligibility index for Mandarin Chinese, Speech Communication, 2016.
4. Liu et al., A new data-driven band-weighting function for predicting the intelligibility of noise-suppressed speech, APSIPA, 2017.

Performance comparison:

1. Van Kuyk et al., An evaluation of intrusive instrumental intelligibility metrics, IEEE/ACM TASLP, 2018.
2. Tang et al., Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech, Computer Speech & Language, 2016.
3. Falk et al., Objective quality and intelligibility prediction for users of assistive listening devices, IEEE Signal Process Mag. 2015.

Interspeech 2016, special session 'Intelligibility under the Microscope'

Table of contents

2.2.1. Language effect

Different languages have different acoustic and lingustic features, e.g., for Manarin and English:

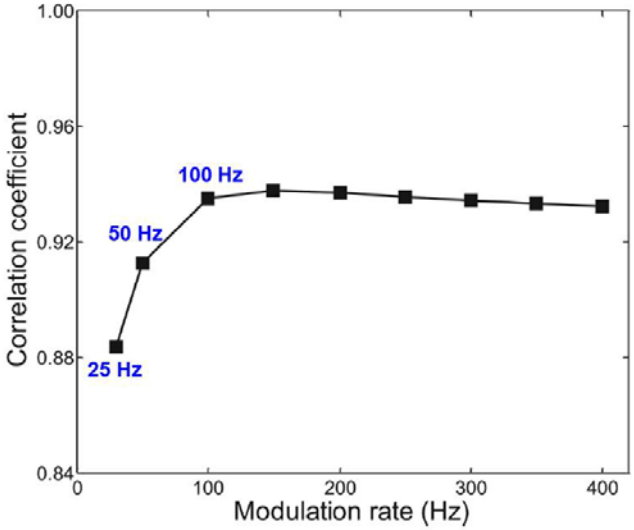
1) Tone contour contribution: lexical meaning?
2) Syllable structure: mono-syllabic in Mandarin vs. multi-syllabic in English (e.g., 'ba' vs. 'manipulate');
3) Numbers of vowels/consonants: 35/21 of vowels/consonants in Mandarin vs. 20/32 in English (Chen et al., JASA, 2013; Fogerty et al., JASA, 2009); and
4) Proportion of vowel segments: 66% in Mandarin vs. 45% in English (Chen et al., JASA, 2013; Fogerty et al., JASA, 2009).

Envelope modulation frequency difference

- For English, a low modulation frequency (0–12.5 Hz) was found to be sufficient for speech intelligibility (e.g., Drullman et al., 1994).
- For a tonal language (e.g., Mandarin), using a high modulation frequency (f_{cut}) (e.g., 100 Hz) is beneficial for intelligibility prediction.

Mandarin sentence

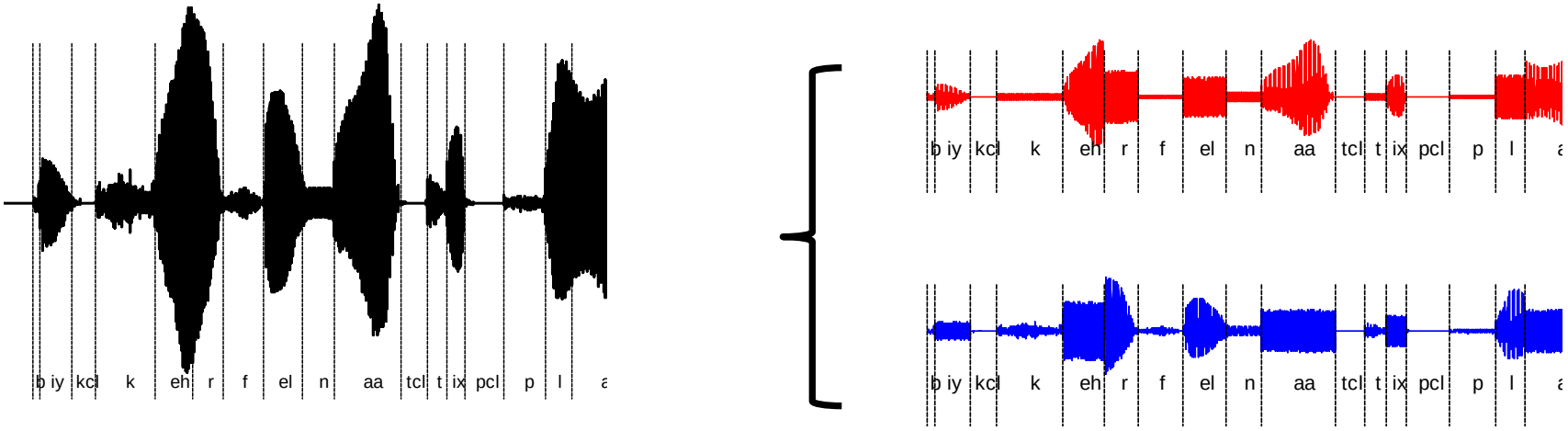
NCM-based prediction



Ma et al., JASA, 2009; Chen et al., JASA, 2011

2.2.2. Segmental contribution: Vowel importance

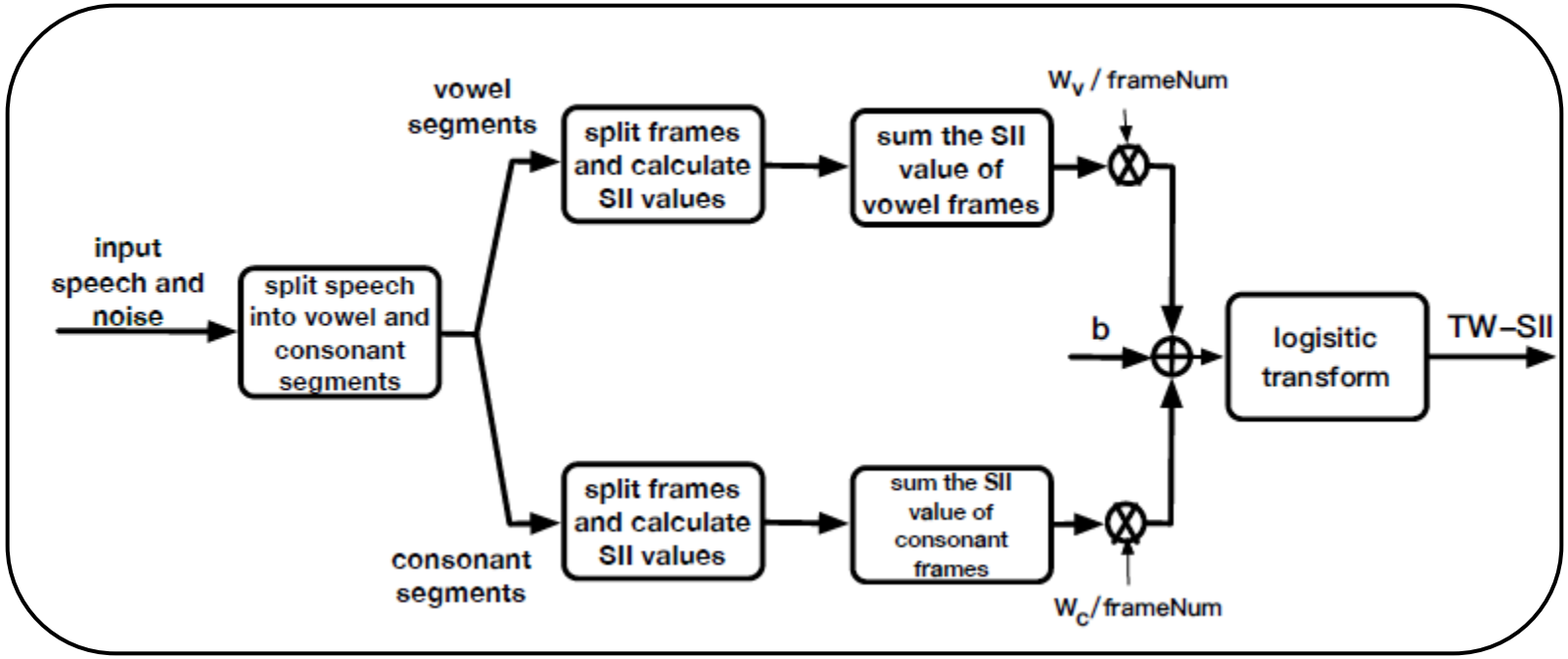
- A remarkable (2:1) advantage of vowels (Vs) vs. consonants (Cs) for English sentence perception (Cole et al., 1996; Kewley-Port et al., 2007).
  - 87.4% (V-only sentences, with Cs replaced by white noise)
  - 46.6% (C-only sentences, with Vs replaced by white noise)



- A much higher (3:1) advantage of vowels vs. consonants in Mandarin (Chen et al., 2013).

Segmental contribution: Vowel importance

- Using a time-weighted function that accounts for the relative perceptual importance of vowels and consonants in speech intelligibility.



Song et al., A time-weighted method for predicting the intelligibility of speech in the presence of interfering sounds, ICASSP, 2018.

2.2.3. Customized model for individual listener

- Most intelligibility indices were designed for evaluating speech perception of normal-hearing listeners.
  - Hearing loss may cause reduced spectral discrimination, hearing dynamic range, etc., negatively affecting speech perception performance.
  - For listeners using hearing aids or cochlear implants (CI), the design of intelligibility index should involve those patient-specific hearing loss factors.
    - Examples include HASPI, HA- and CI- versions of SRMR, etc.

HA: hearing aids
CI: cochlear implants

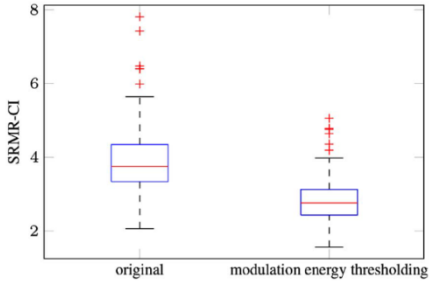
Kates et al., Ear and Hearing, 2018; Tiago, IEEE Sig. Proc. Mag., 2015

SRMR-CI metric for cochlear implant users

- First, the gammatone filter bank was replaced by the filter bank used in the speech coding strategy of the CI devices.
- Second, speech content variability was reduced by means of a modulation spectrum thresholding scheme.

$$\texttt{clamp}(x, a, b) = \begin{cases} a & \text{if } x \leq a \\ x & \text{if } a < x \leq b \\ b & \text{if } x > b \end{cases}$$

Limited in the range [a, b]



- Finally, to model the reduced sensitivity of CI listeners, the center frequency range (4–128 Hz) of the 8 modulation filters of the original SRMR metric was reduced to 4–30 Hz.
- The SRMR-CI has been tested as a correlate of intelligibility for CI users under clean, noisy, reverberant, noise-plus-reverberation, and speech-enhanced conditions.

Santos and Tiago, IEEE/ACM TASLP, 2014

2.2.4. Compensate the misalignment between probe and response waveforms

- Due to speech signal processing, the probe and processed speech signals are not temporally aligned, which may affect the performance of intelligibility prediction.
  - Kates and Arehart (2005) stated that for greatest accuracy, unbiasing techniques that temporally align the input and output sequences should be used when computing the magnitude-squared coherence (MSC).

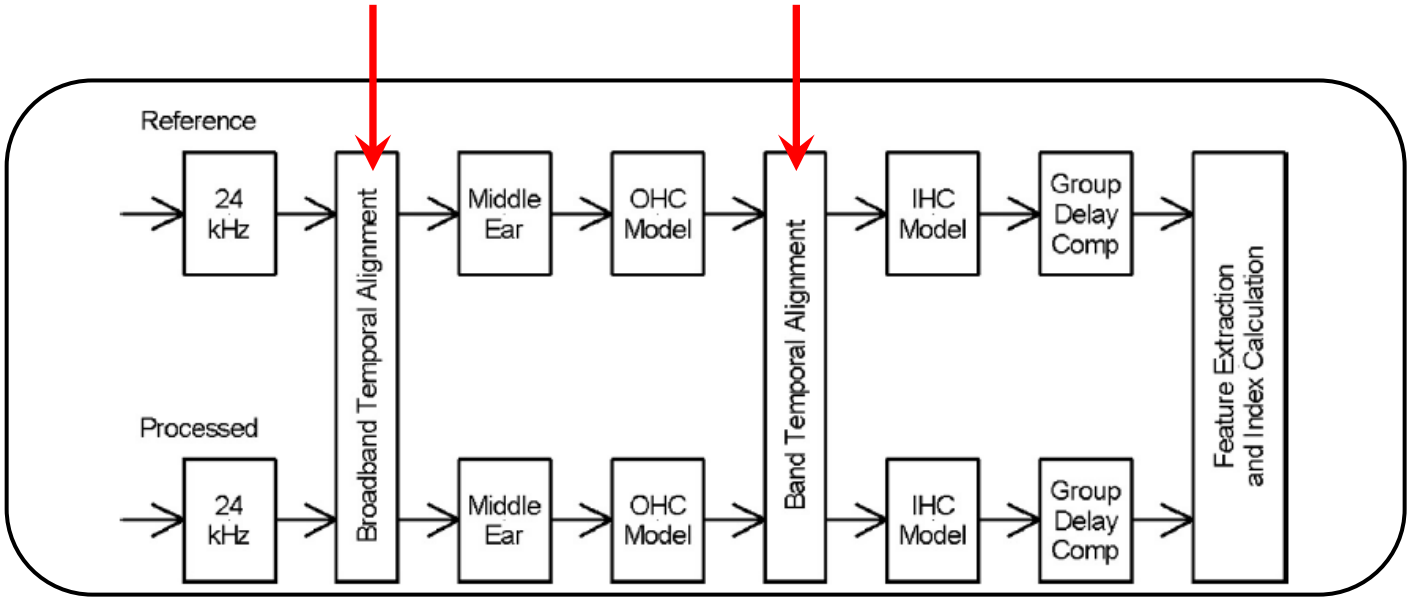
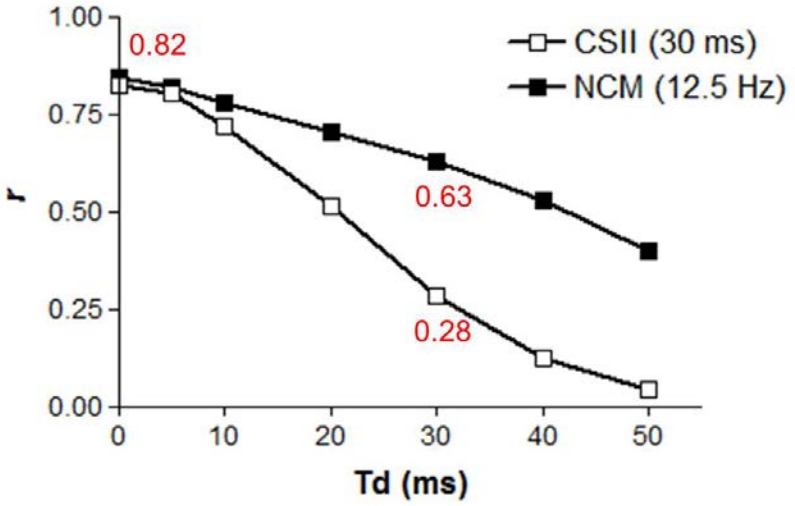


Fig. 1. Block diagram showing the reference and processed signal comparison.

Effect of temporal misalignment on intelligibility prediction

Intelligibility data 1: noise-suppressed speech



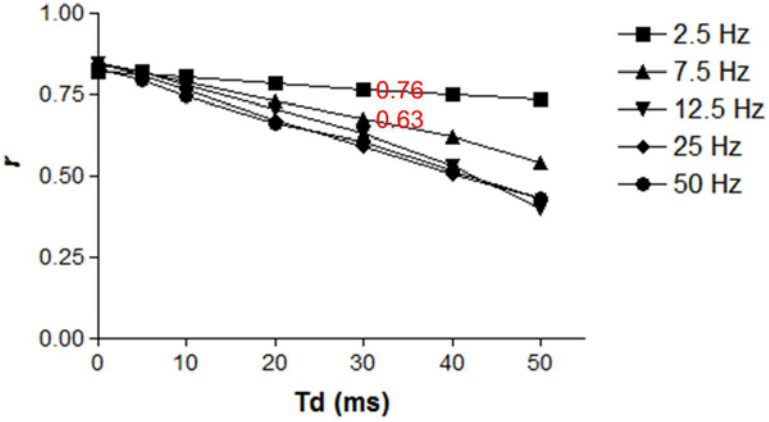
Td: temporal misalignment between probe and processed waveforms

- The envelope-based NCM was more resistant to temporal misalignment than the CSII, which was computed using short-time spectral detail.

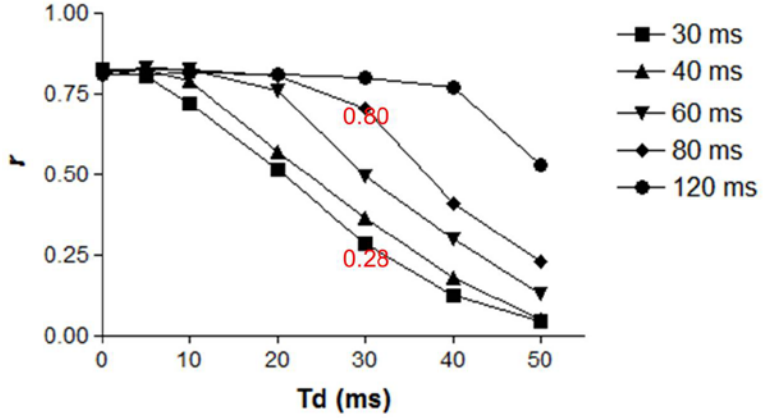
Chen et al., Compensating the effect of signal-processing delay on objective intelligibility prediction, 23rd International Conference on Digital Signal Processing (DSP), 2018

Compensate the effect of temporal misalignment on intelligibility prediction

NCM:
modify f_{cut} for temporal envelope cue



CSII:
modify the segmentation duration T



- The negative influence of signal-processing delay on objective intelligibility prediction could be compensated by employing
  - a small frequency range of envelope modulation in the temporal-envelope, or
  - a long segmentation duration in the spectral-detail-based indices

Chen et al., Compensating the effect of signal-processing delay on objective intelligibility prediction, 23rd International Conference on Digital Signal Processing (DSP), 2018

More … (3)

Segmental effect:

1. Wang et al., Assessing the segmental contribution to the non-intrusive intelligibility prediction of noise-suppressed speech, IWAENC, 2016.
2. Chen et al., Contributions of cochlea-scaled entropy and consonant-vowel boundaries to prediction of speech intelligibility in noise, JASA, 2012

Language effect:

1. Chen et al., Predicting the intelligibility of vocoded and wideband Mandarin Chinese, JASA, 2011.
2. Chen et al., Non-intrusive intelligibility prediction for Mandarin speech in noise, TENCON 2013.
3. Chen et al., Predicting the intelligibility of vocoded speech, Ear and Hearing, 2011.
4. Li et al., Comparative intelligibility investigation of single-channel noise reduction algorithms for Chinese, Japanese, and English, JASA, 2011.

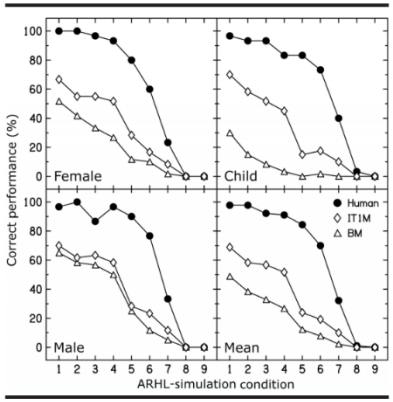
Table of contents

3.1. Automatic speech recognition based

Fontan et al., 'Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss,' J Speech Lang Hear Res, 2017.

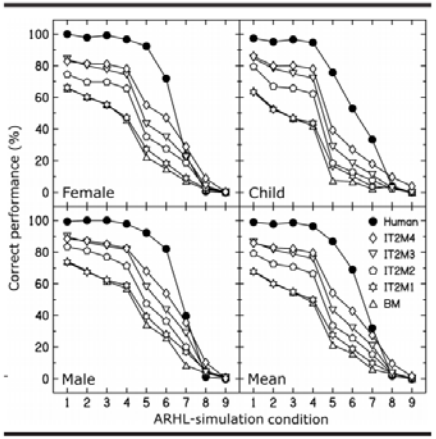
- Speech Material: 60 words (6 lists of 10 words) and 60 sentences (3 lists of 20 sentences) from female child, and male speakers.
- Simulation of age-related hearing loss (ARHL): 9 levels of hearing-loss severity.
- In addition to ASR experiments, 60 listeners completed two intelligibility tests (word and sentence).

Word intelligibility



IT1M, BM:
ASR with 2 language models

Sentence intelligibility



Correlation between human intelligibility scores and ASR scores

Model	Speaker			
	Mean	Male	Female	Child
BM	.94 (.000)	.93 (.000)	.97 (.002)	.71 (.032)
IT1M1	.97 (.000)	.95 (.000)	.99 (.000)	.93 (.000)**

Model	Speaker			
	Mean	Male	Female	Child
BM	.94 (.000)	.95 (.000)	.94 (.000)	.90 (.001)
IT2M1	.96 (.000)**	.97 (.000)	.96 (.000)*	.94 (.000)**
IT2M2	.97 (.000)**	.97 (.000)**	.97 (.000)	.95 (.000)***
IT2M3	.98 (.000)**	.98 (.000)*	.97 (.000)*	.97 (.000)***
IT2M4	.99 (.000)***	.99 (.000)*	.98 (.000)	.96 (.000)***

3.2. Machine learning based

Sharma, et. al., "A data-driven non-intrusive measure of speech quality and intelligibility", Speech Communication, 2016

Dataset:

- Training set:
  8 utterances by each of the speakers (168 speakers in total) from TIMIT database.
- Test set:
  8 utterances by each of the speakers (another 168 speakers) from TIMIT database.
- 19 noise signals from the NATO noise database are added to the training set and test set, with SNRs from −24 to 30 dB in 3 dB steps.

Speech feature extraction:

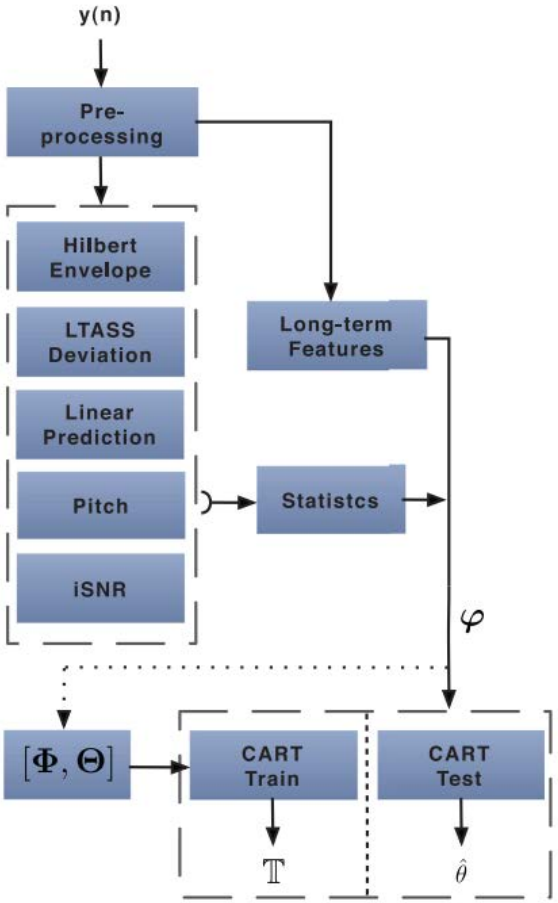
- Short-term features: the statistical features and their first-order differences (as shown in the table)
- Long-term features: the magnitude spectrum of the speech signal over the entire utterance

NISA: non-intrusive speech assessment

		LCQA		LCIA		NISA	
Per-frame feature	Feature Id	x	$\Delta(x)$	x	$\Delta(x)$	x	$\Delta(x)$
LPC spectral flatness	ϕ_1	x	x	x	x	x	x
LPC spectral dynamics	ϕ_2	x		x		x	
LPC spectral centroid	ϕ_3	x	x	x	x	x	x
LPC excitation variance	ϕ_4	x	x	x	x	x	x
Speech variance	ϕ_5	x	x	x	x	x	x
Pitch (LPC)	ϕ_6	x	x				
Pitch (PEFAC)	ϕ_7					x	x
Zero crossing rate	ϕ_8					x	x
iSNR	ϕ_9			x	x	x	x
Hilbert envelope variance	ϕ_{10}					x	x
Hilbert envelope range	ϕ_{11}					x	x
Spectral flatness (PLD)	ϕ_{12}					x	x
Spectral dynamics (PLD)	ϕ_{13}					x	x
Spectral centroid (PLD)	ϕ_{14}					x	x

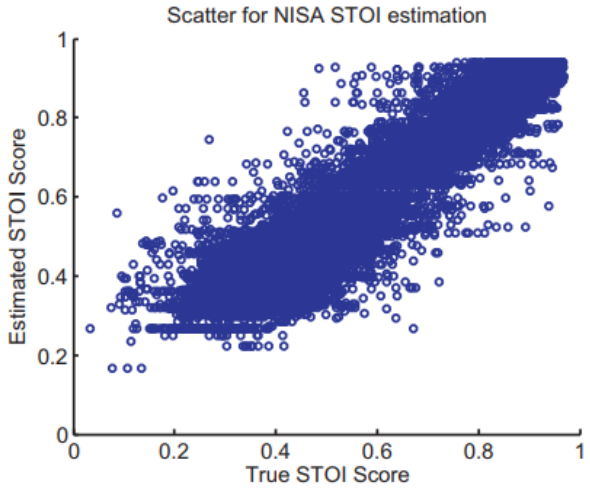
Model training:

- Classification and Regression Tree (CART) is trained using a minimum mean square error (MSE) criterion, with 10-fold cross-validation of the training data.



Metrics:(between the predictions and real signals)

- Spearman Correlation Coefficient (SCC)
- Root Mean Square Error (RMSE)
- Bin error
- Two class hit rate (TCHR)



			Bin error				
Method	SCC	RMSE	0.05	0.10	0.15	0.20	TCHR
NISA	**0.95**	**0.08**	**64.0**	**85.4**	**93.3**	**97.0**	**94.7**
LCIA	0.91	0.18	26.6	45.4	61.3	72.9	80.9
LCQA	0.65	0.19	9.4	19.9	35.9	59.8	62.6

3.2. Machine learning based

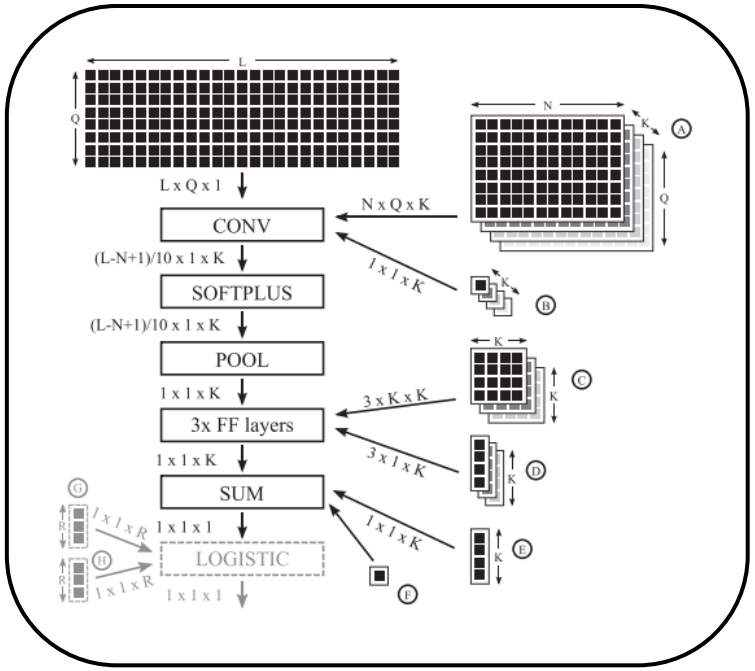
Andersen et. al., "Nonintrusive speech intelligibility prediction using convolutional neural networks," IEEE/ACM TASLP, 2018.

Dataset:

- 4 listening experiments datasets (D1, D2, D3, D4) including 525 conditions (i.e., different languages, genders, speakers, words and SNRs)
- The combined dataset is split into training set, test set and validation set with the 4:1:1 ratio. Each set is with the similar composition of different conditions.

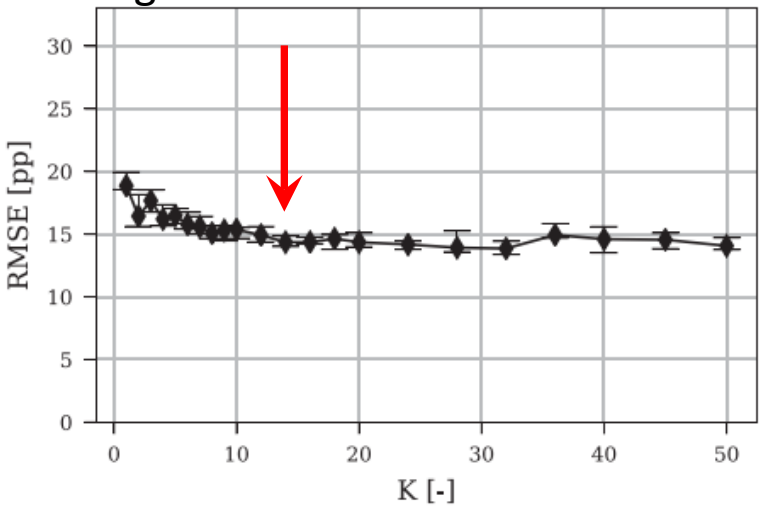
Speech feature extraction:

- Speech signals are resampled to 10 kHz, and periods without speech are removed by use of an ideal VAD.
- Then signals are analyzed with a short-time Discrete Fourier Transformation (DFT) as the speech feature. This is done in 256-sample Hann-windowed segments which are zero-padded to 512 samples.



CNN-based non-intrusive intelligibility prediction

Figure 1



the number of convolution kernels

Figure 2

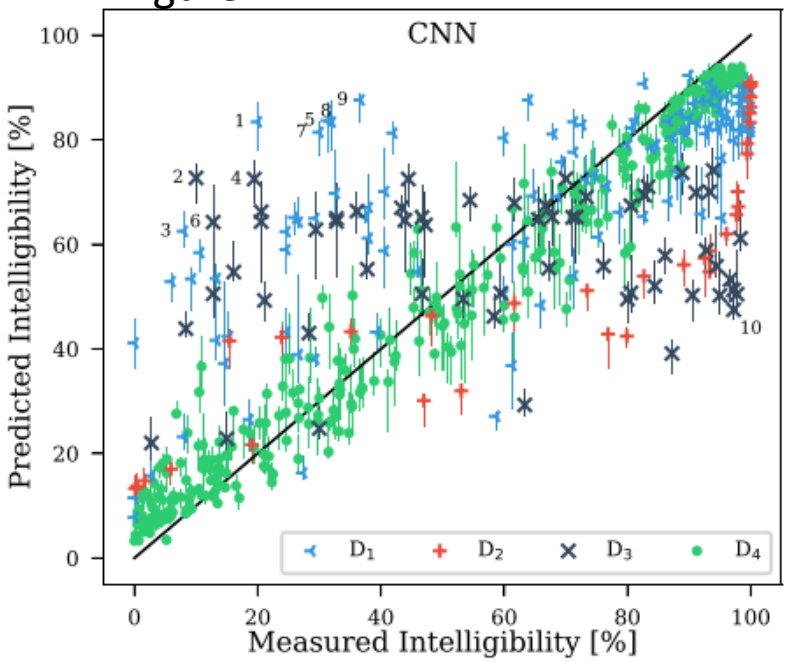


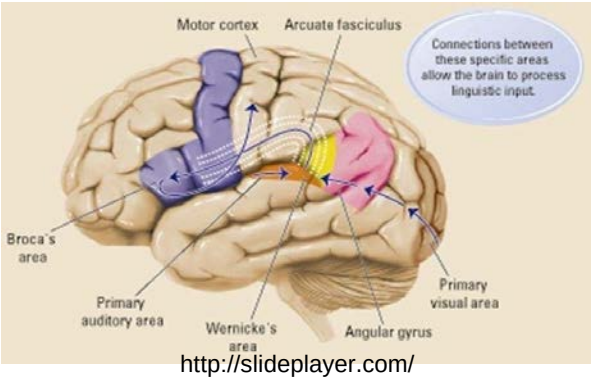
TABLE III
PERFORMANCE METRICS FOR THE FIVE SIP ALGORITHMS

SIP algorithm	RMSE	Kendall's Tau
CNN	17.69 pp	0.667
STOI	18.94 pp	0.658
ESTOI	17.11 pp	0.692
NI-STOI	19.90 pp	0.629
SRMR	32.77 pp	0.281

3.3. Brain neural activity based

Advantages:

- Brain activity [electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS)] based measures of speech intelligibility are objective, and can be performed automatically in response to natural running speech.
- Understanding the neural processes associated with speech perception may help dissociate mechanisms in populations who have similar behavioral performance speech perception tests, but differ in the underlying source of dysfunction.
- fNIRS is suitable for use with both adult and pediatric listeners, and particularly for CI recipients.

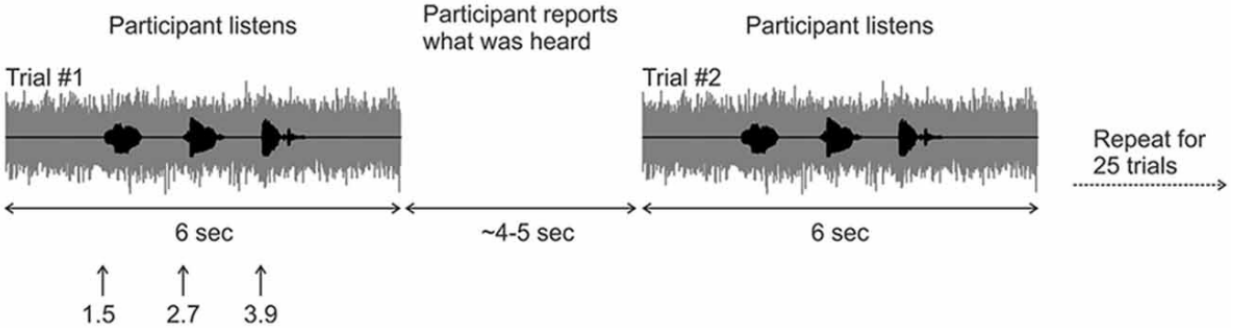


http://slideplayer.com/

3.3.1. EEG based speech intelligibility prediction

Dimitrijevic et al., Cortical alpha oscillations predict speech intelligibility. Frontiers in Human Neuroscience, 2017.

- Subjects: 14 normal-hearing adult participants.
- Stimuli: an introductory phrase “The numbers” and monosyllabic digits 0–9 excluding the disyllabic 7, where the “0” was pronounced “Oh” (/oω/).
- Behavioral experiments: The triplet digits were presented in noise in successive trials to estimate speech reception threshold (STR).
  - The procedure of estimating STR:
    - Trials following a correct response (all three digits) reduced the SNR by 2 dB (noise constant, digit amplitudes reduced).
    - Incorrect responses were followed by an increased SNR, also by 2 dB.
    - The average SNR over the last 11 trials was the SRT.



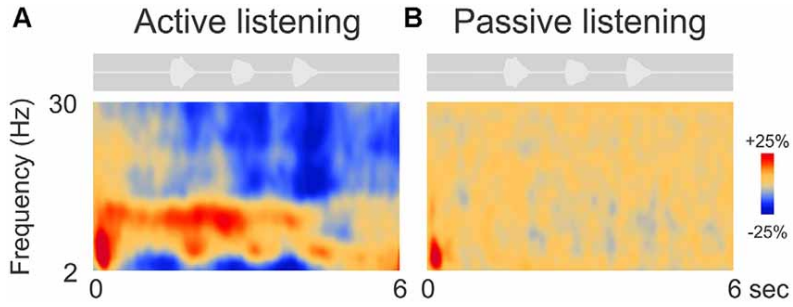
- Electrophysiological experiments:
  - “passive” listening condition: participants were instructed to ignore any sounds while they watched a closed caption and silent movie of their choice.
  - “active” listening condition: participants fixated a white cross on a computer screen and repeated verbally all the digits presented.

3.3.1. EEG based speech intelligibility prediction

Dimitrijevic et al., Cortical alpha oscillations predict speech intelligibility. Frontiers in Human Neuroscience, 2017.

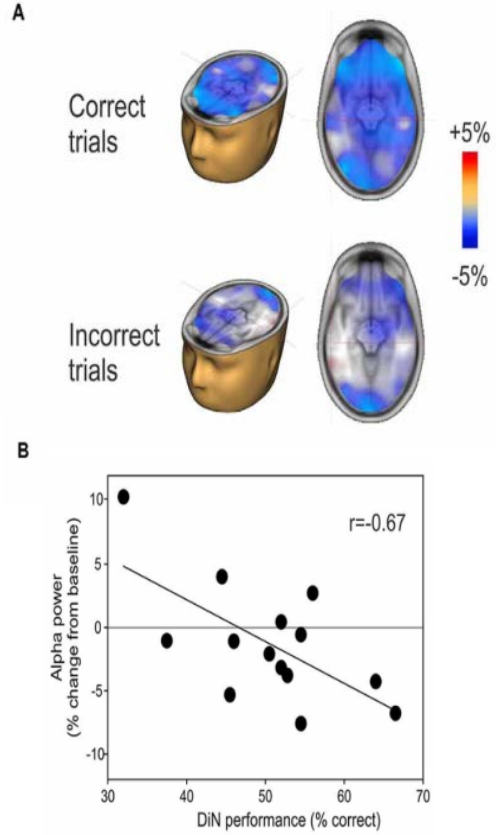
Effects of attention:

- **red**; event-related synchronization, ERS
- **blue**; event-related desynchronization, ERD
- The active condition is characterized by more oscillatory power in the alpha (8–12 Hz), beta (15–30 Hz) delta/theta (2–6 Hz) bands than in the passive listening condition.



Alpha power and digit identification

- Panel A shows when the digits were identified correctly, the alpha **ERD** was of greater magnitude in left temporal regions, and the alpha **ERS** showed no consistent difference.
- Panel B shows a significant negative correlation between alpha **ERD** and digit-in-noise performance.



3.3.2. fNIRS based speech intelligibility prediction

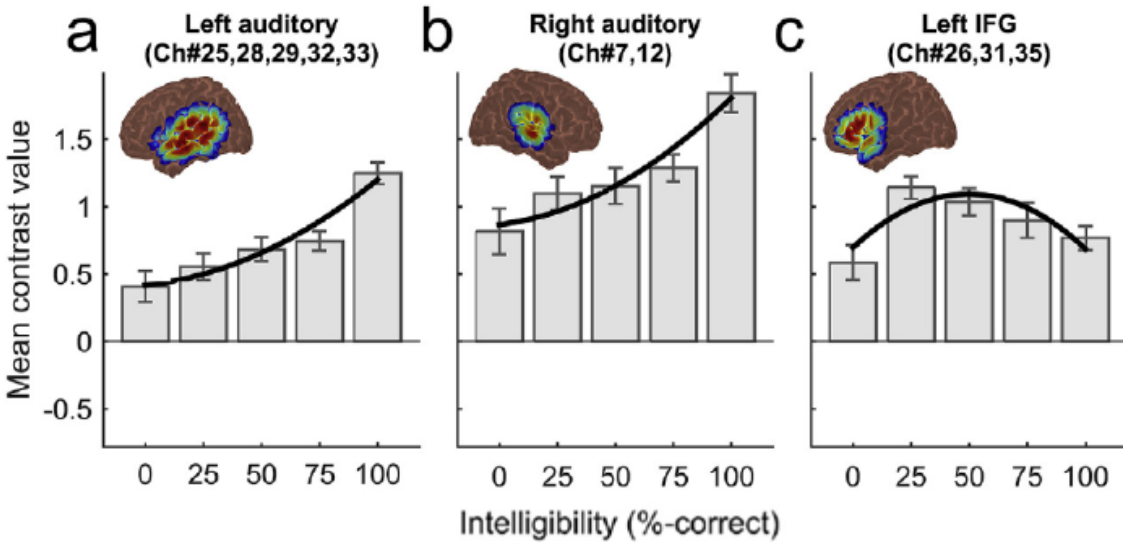
Lawrence t al., Cortical correlates of speech intelligibility measured using functional near-infrared spectroscopy (fNIRS), Hearing Research, 2018.

- Subjects: 23 healthy adult
- Speech stimuli: 8-channel noise-vocoded Bamford-Kowal-Bench (BKB) sentences
  - Envelope exponents of 0.000, 0.149, 0.212, 0.297 and 1.000 was used, chosen to target group-mean intelligibility levels of 0, 25, 50, 75 and 100% keywords correct, respectively.
  - These 5 stimulation conditions are referred to as S0, S25, S50, S75 and S100.
- Behavioral test of speech intelligibility:
  - before and after the main fNIRS task
  - respond verbally by attempting to repeat what they had heard
- Main fNIRS task:
  - Participants were presented with 20 sentences per stimulation condition.
  - After each stimulus ended, a probe word appeared on the display. Participants were required to indicate by a button press whether the probe word had appeared in the sentence just heard.

3.3.2. fNIRS based speech intelligibility prediction

Lawrence t al., Cortical correlates of speech intelligibility measured using functional near-infrared spectroscopy (fNIRS), Hearing Research, 2018.

- Results in main fNIRS task : Response profiles in regions-of-interest
  - Panels a and b show in left and right auditory regions, activation increased monotonically as intelligibility improved.
  - Panel c shows in the left inferior frontal gyrus, activation increases monotonically as intelligibility was reduced from the S100 condition down to the S25 condition.



More … (4)

1. Lesenfants et al., Predicting individual speech intelligibility from the cortical tracking of acoustic-and phonetic-level speech representations, Hearing Research, 2019.
2. Iotzov et al., EEG can predict speech intelligibility, Journal of Neural Engineering, 2019.
3. Etard et al., Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise, Journal of Neuroscience, 2019.
4. Defenderfer et al., Investigating the role of temporal lobe activation in speech perception accuracy with normal hearing adults: An event-related fNIRS study, Neuropsychologia, 2017.
5. Olds et al., Cortical activation patterns correlate with speech understanding after cochlear implantation, Ear and Hearing, 2016.
6. Pollonini et al., Auditory cortex activation to natural speech and simulated cochlear implant speech measured with functional near-infrared spectroscopy, Hearing Research, 2014.

Table of contents

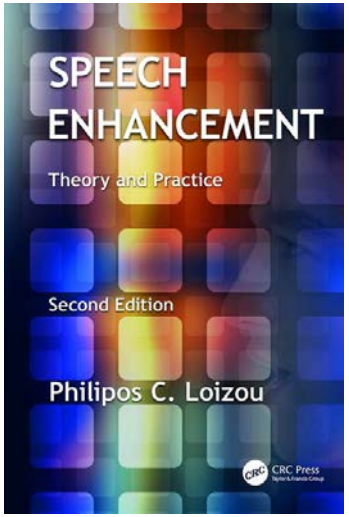
Summary

- Objective intelligibility evaluation is a very useful tool for speech studies (speech perception, enhancement, processing, etc.)
- However, there are still challenges in order to further improve the intelligibility evaluation performance in different application scenarios.
  - Acoustic cues (particularly temporal envelope) for speech perception are the basis for most indices.
  - Nonintrusive intelligibility index is of practical significance.
  - For challenging listening scenarios, significant efforts for optimization/customization are needed for improving performance.
  - The design of evaluation methods is developing, with ASR, machine learning, brain science technologies.
- We still need to increase our multidisciplinary (acoustics, linguistics, psychology, engineering, brain science, etc.) knowledge on speech perception mechanisms.

Access …

Intrusive intelligibility index

- Matlab codes of NCM, MSC, CSII (Loizou, 2007)
  - CD of book <*Speech Enhancement: Theory and Practice*>
- Matlab code of STOI (Taal et al., IEEE/ACM TASLP, 2011)
  - http://siplab.tudelft.nl/
- Matlab code of eSTOI (Jensen and Taal, IEEE/ACM TASLP, 2016)
  - http://kom.aau.dk/~jje/



Non-intrusive intelligibility index

- Matlab code of SRMR (Tiago et al., IEEE/ACM TASLP, 2010)
  - http://musaelab.ca/software/
- Matlab code of ModA (Chen et al., BSPC, 2013)
  - https://eee.sustech.edu.cn/feichen/information.html

All intelligibility data will be provided upon request to fchen@sustech.edu.cn

THANKS
Q & A

INTERSPEECH 2020
OCTOBER 25-29/ SHANGHAI , CHINA
SHANGHAI INTERNATIONAL CONVENTION CENTER