13th Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Tutorial T1

Speech Perception and Enhancement in Cochlear Implants

Fei Chen

Professor, Speech and Physiological Signal Processing Laboratory

Department of Electrical and Electronic Engineering

Southern University of Science and Technology (SUSTech), Shenzhen, China

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

14/12/2021

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

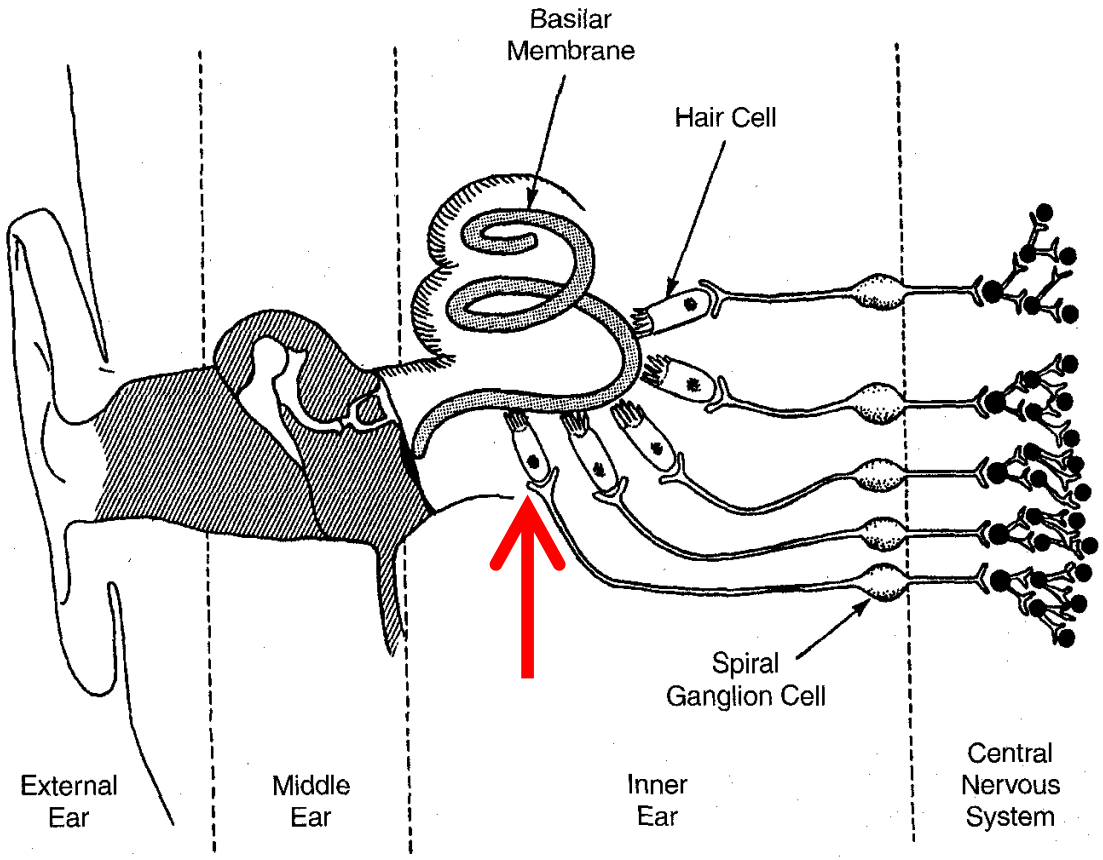
电子与电气工程系 DEPARTMENT OF ELECTRONIC AND ELECTRICAL ENGINEERING

Outline

1. Background
   - Cochlear implants (CIs)
   - Important cues for human speech perception
2. Cochlear implants speech perception
   - Vocoder model for simulating CI speech perception
   - Combined acoustic-electric stimulation
   - Objective intelligibility evaluation for CI speech perception
3. Summary

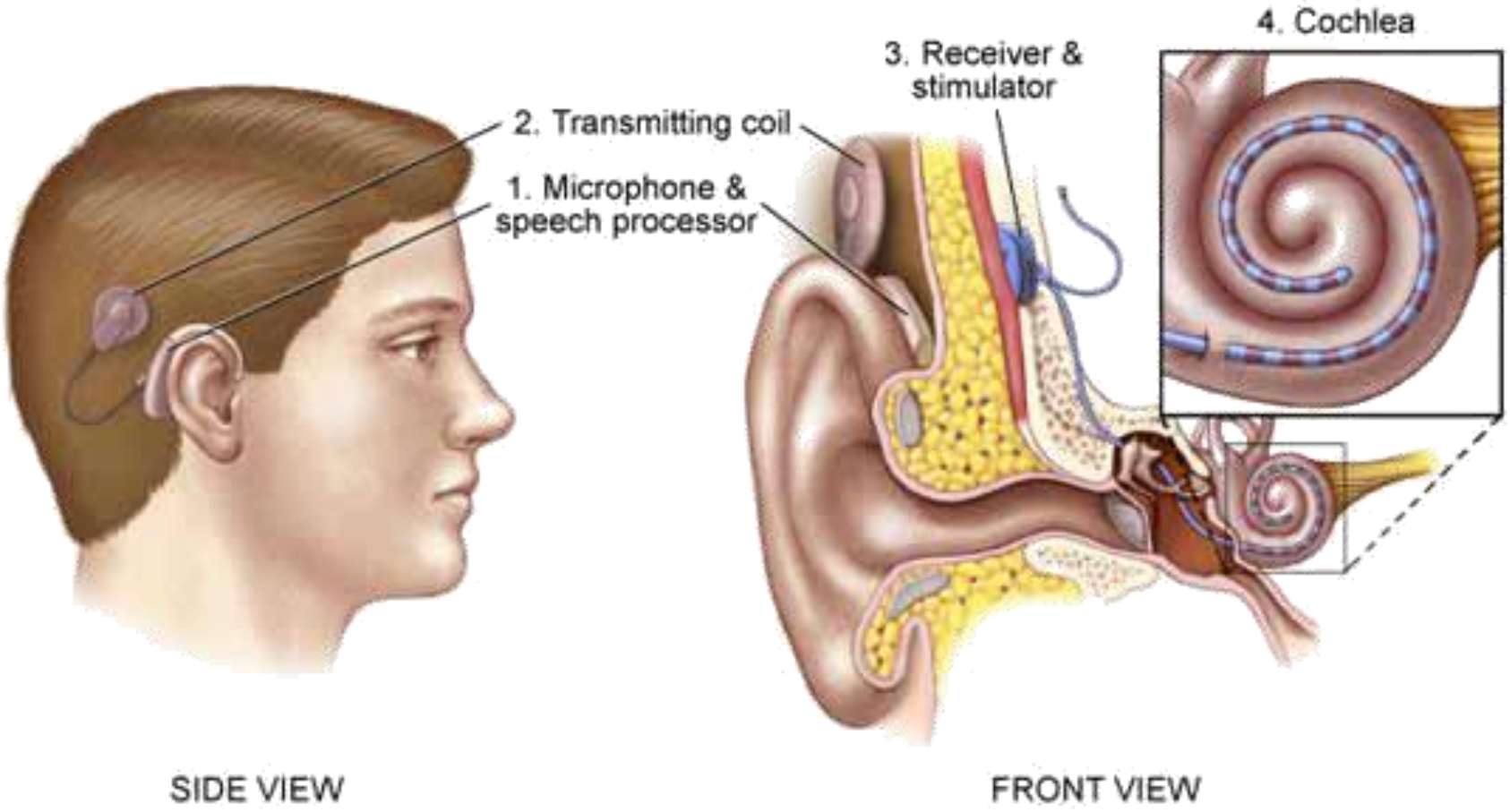
1.1 Cochlear implants (CIs)

- Cochlear implants (CIs): When there is damage to the inner ear (cochlea) or to the nerve pathways from the inner ear to the brain.

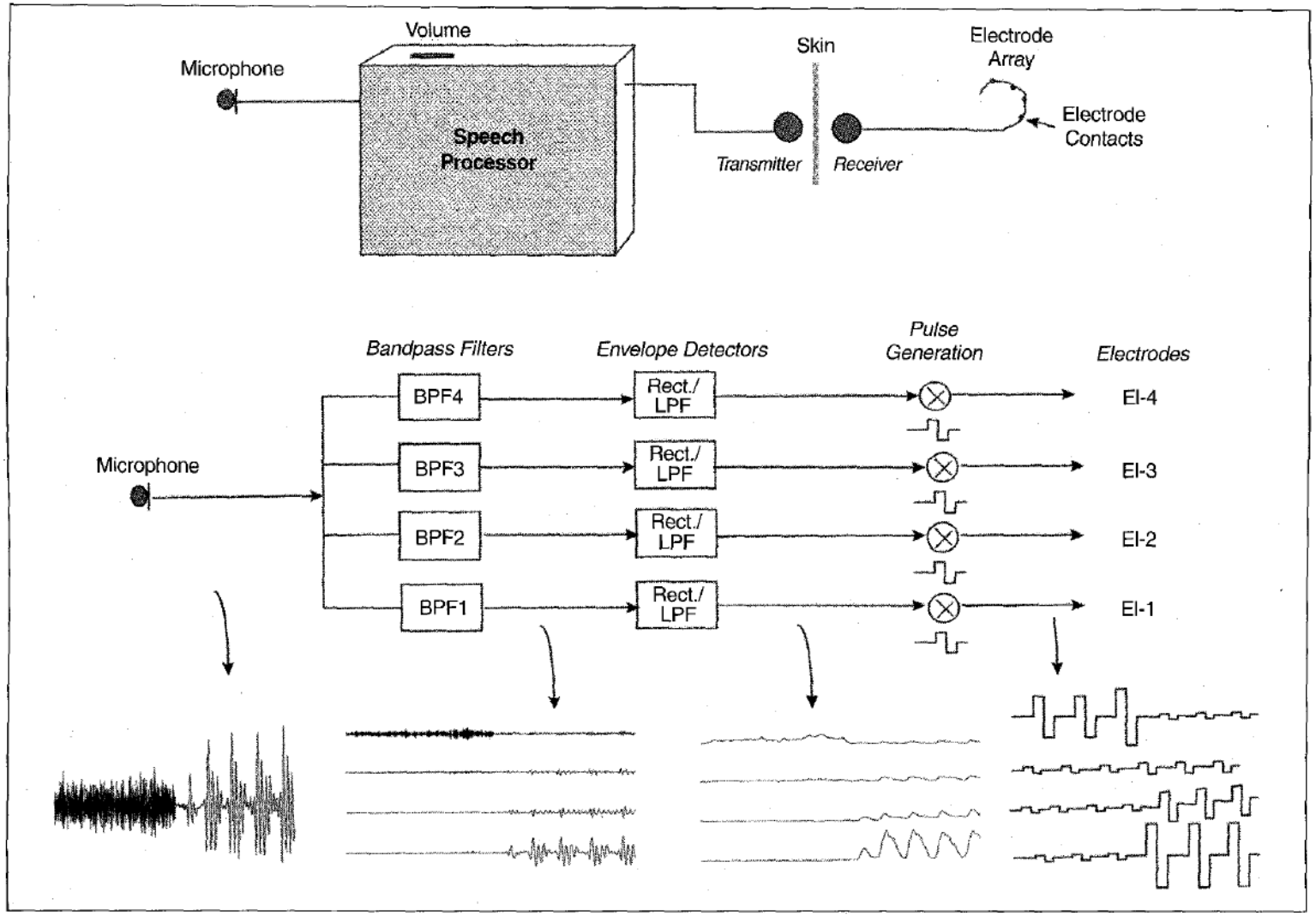


Loizou, IEEE EMBM, 1999

Cochlear implants (cont.)



Block diagram of a CI system

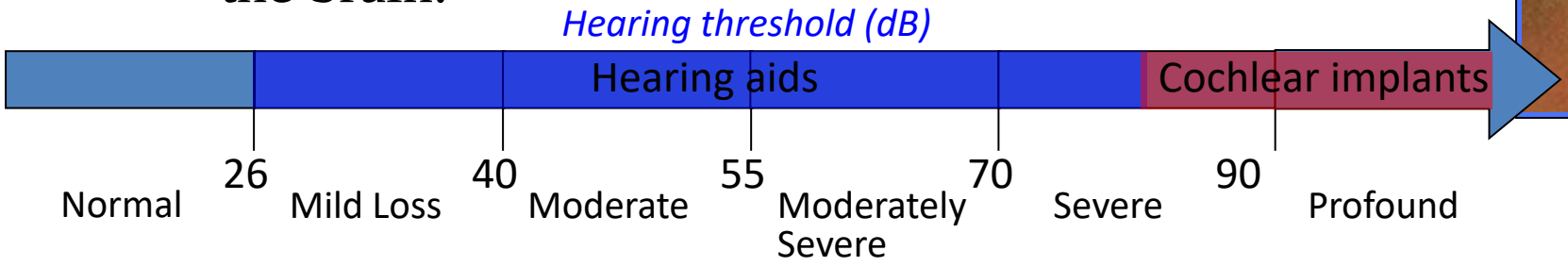


Loizou, IEEE EMBM, 1999

Assistive hearing devices: CIs and HAs

1. Hearing aids (HAs) and conductive hearing loss
   - When sound signal is not conducted efficiently through the outer and middle ears.
2. Cochlear implants (CIs) and sensorineural hearing loss
   - When there is damage to the inner ear (cochlea) or to the nerve pathways from the inner ear to the brain.

Images: drkirtane.com



CI challenges

1) Some users of CIs still do not have high levels of speech perception.
   - There is a large performance variance among implanted users.
2) Understanding of tonal language
   - Most of present CI speech processors are designed for non-tonal language (e.g., English)
3) Music appreciation
4) Speech perception in noise for CI patients
5) Spatial hearing or sound localization is absent for users of unilateral CIs.

Outline

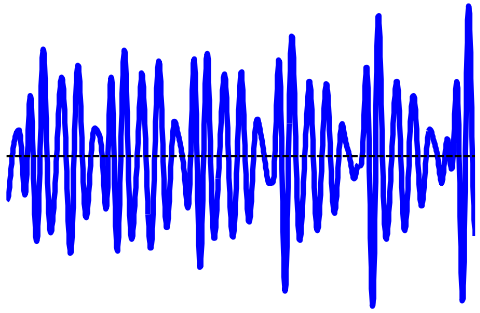
1. Background
   - Cochlear implants (CIs)
   - Important cues for human speech perception
2. Cochlear implants speech perception
   - Vocoder model for simulating CI speech perception
   - Combined acoustic-electric stimulation
   - Objective intelligibility evaluation for CI speech perception
3. Summary

1.2 Important cues for human speech perception

- Temporal cues (temporal envelope, and phase or fine structure)
- Segmental (vowels and consonants)
- Fundamental frequency (F0), etc.

1.2.1 Temporal cues: Envelope and fine structure

- Rosen's definition (1992)
  - Envelope: speech amplitude fluctuation at rates 2-50 Hz
  - Periodicity: 50-500 Hz
  - Fine structure: 600-10 kHz



- Hilbert transform

  Signal= envelope × fine structure

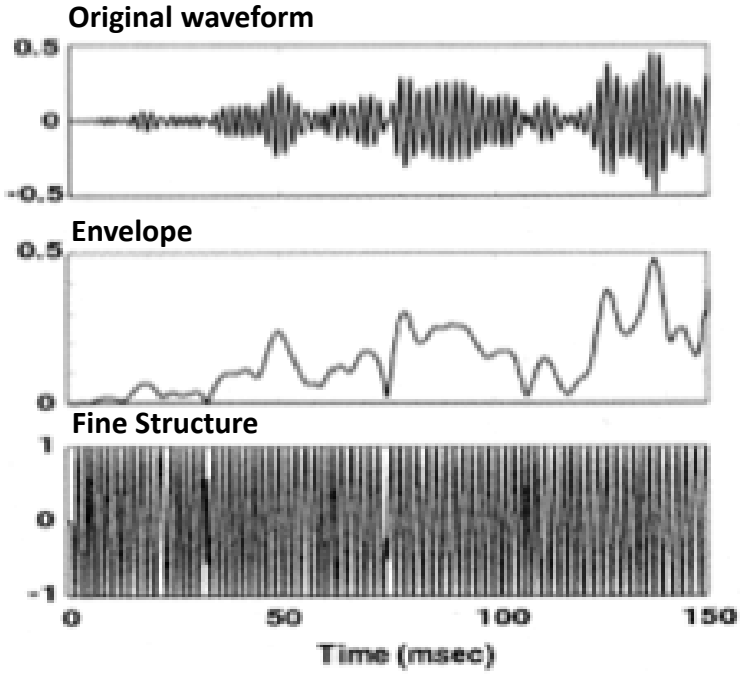


Image from B.S. Wilson et al. (2005), "Two new directions in speech processor design for cochlear implants," Ear Hear.

Temporal envelope for human speech perception

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.

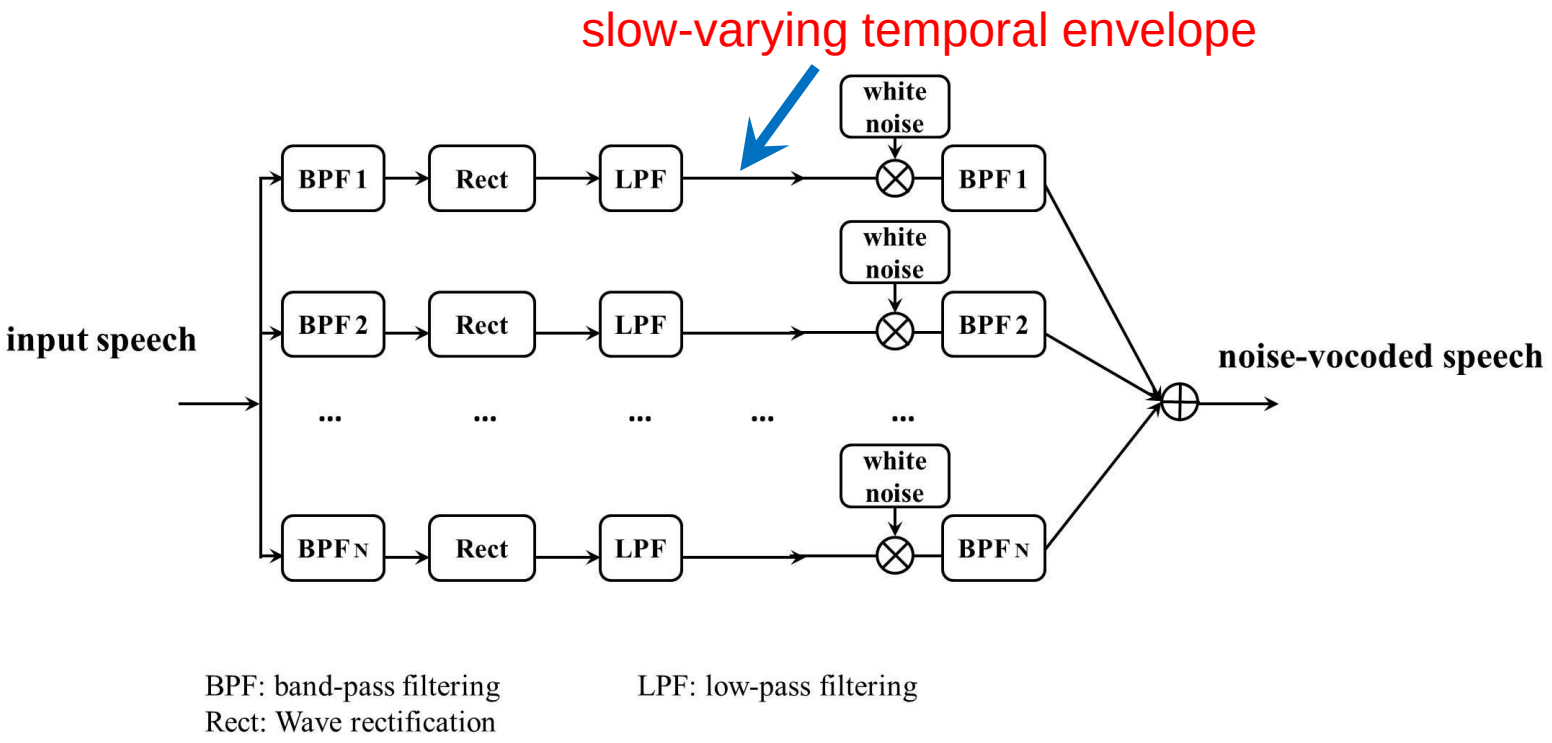
under conditions of reduced spectral cues, slowly varying temporal information (<50 Hz) can yield relatively high speech recognition performance. This result is consistent with the observation of poor speech discrimination in children who have central processing disorders that disrupt temporal processing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

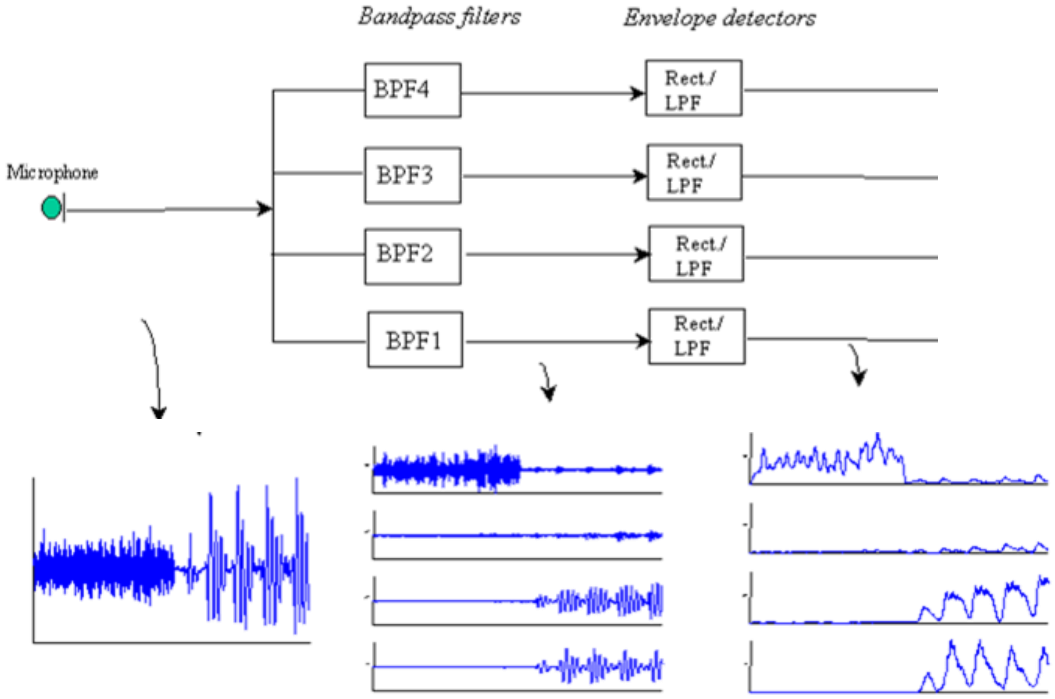
Vocoder model for envelope-based speech synthesis

- Noise vocoder and noise-vocoded speech



"Effects of noise suppression and envelope dynamic range compression to the intelligibility of vocoded sentences for a tonal language," JASA, 2017

Vocoder model for envelope-based speech synthesis (cont.)



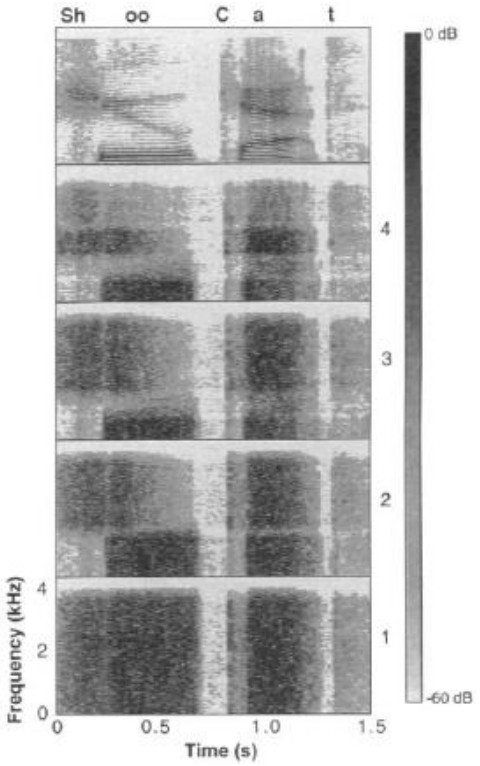
# of bands	Wideband	Noise vocoded
N=1		
2		
4		
8		

'A boy falls from the window'

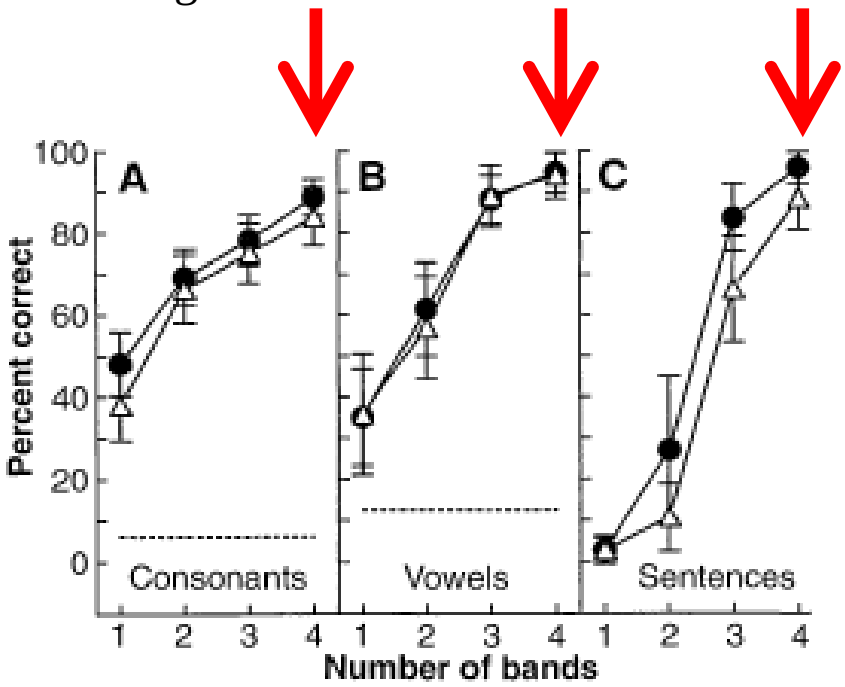
Loizou, IEEE EMBM, 1999

Speech perception with envelope cues

- In vocoder model, spectral information or fast-varying oscillation was largely removed, and only temporal slow-varying envelope information was preserved.



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



Shannon et al., Science, 1995

F0 contour for Mandarin speech perception

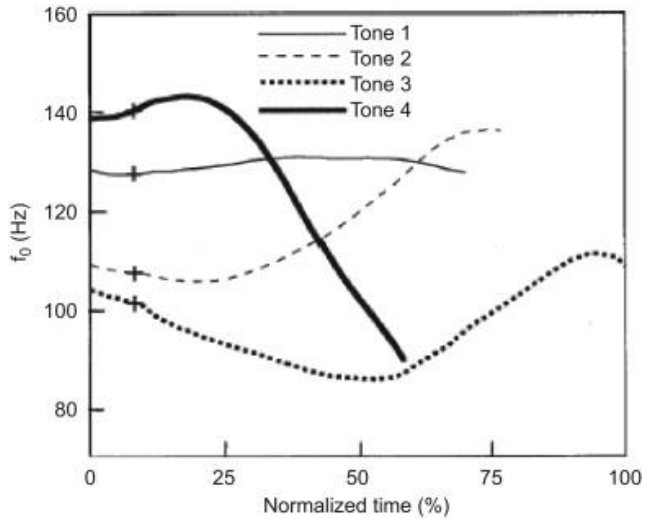


Image: Xu, 1997

Important for lexical tone identification

"zhu"

Tone 1: pig

Tone 2: bamboo

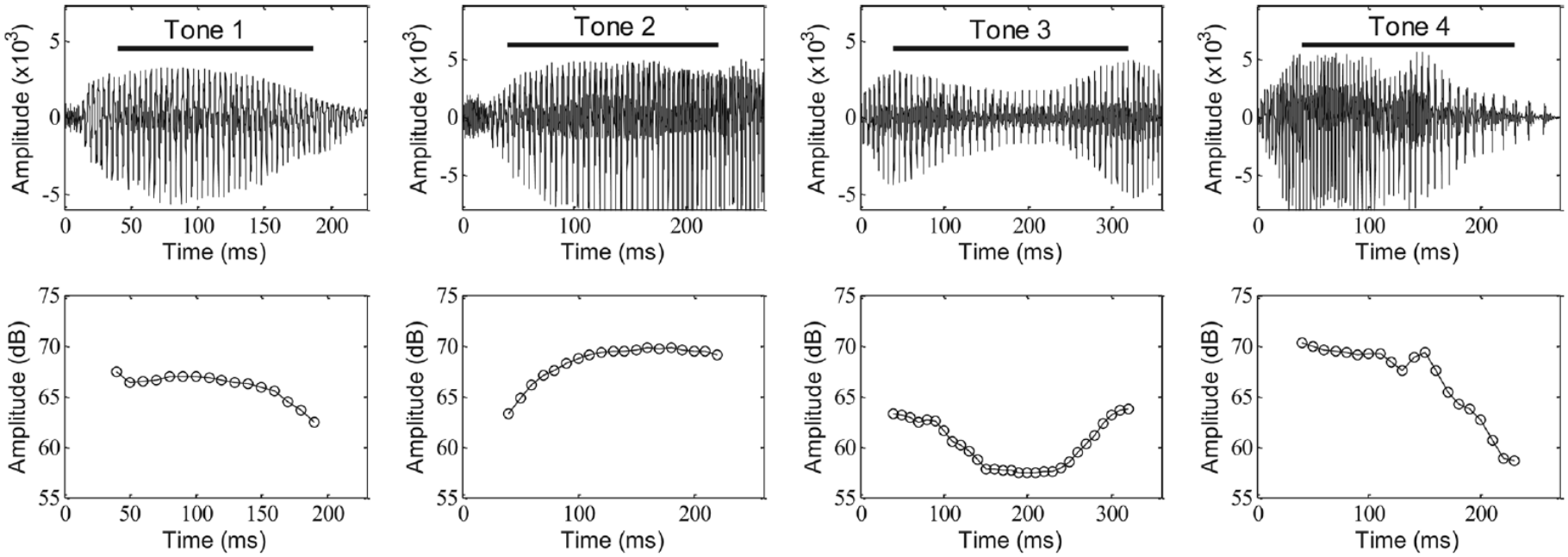
Tone 3: stew

Tone 4: pillar

Images: Yuen, et al. MAPPID-N

Envelope in tonal language: Tonal envelope contour

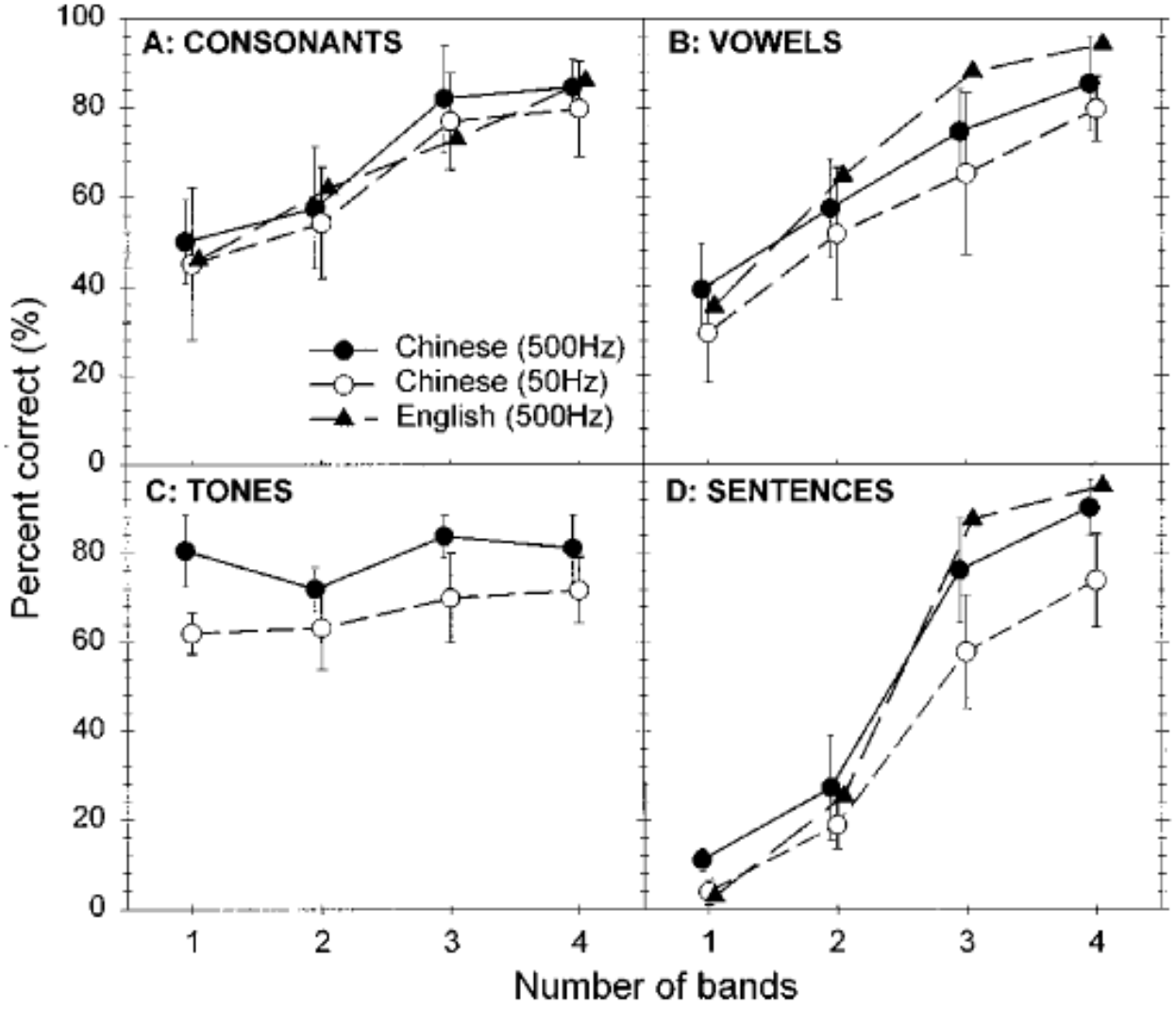
4 tones in Mandarin Chinese



- Implication:
  - Tonal envelope contour can assist Mandarin tone identification.

"Predicting the intelligibility of vocoded and wideband Mandarin Chinese," JASA, 2011

Mandarin speech perception with envelope cues



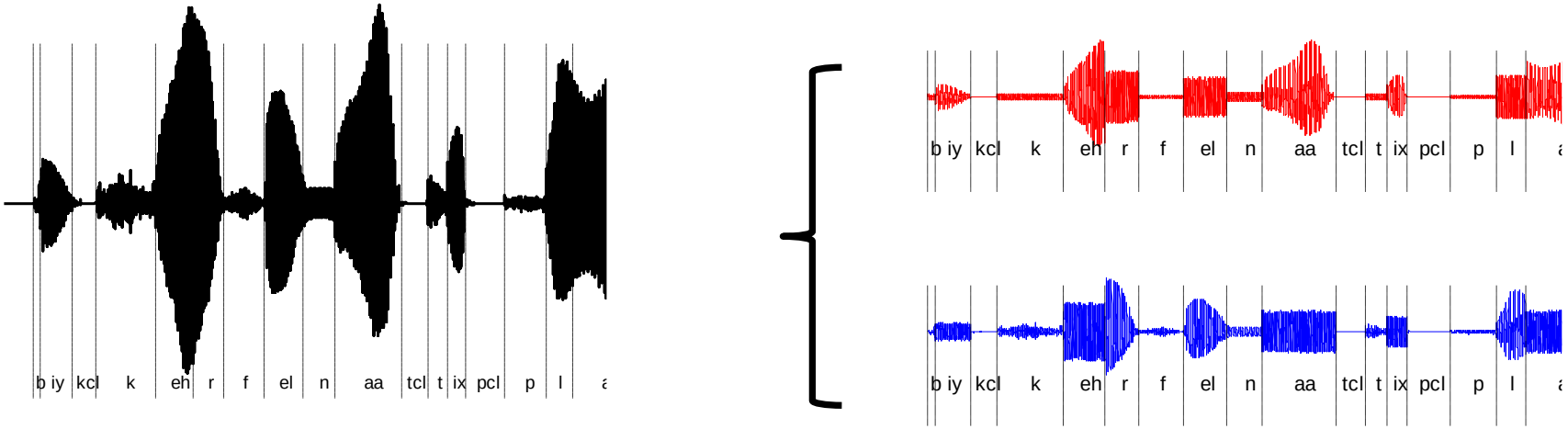
Low-pass filter for envelope

500 Hz (●) and at 50 Hz (○)

Fu et al., JASA, 1998

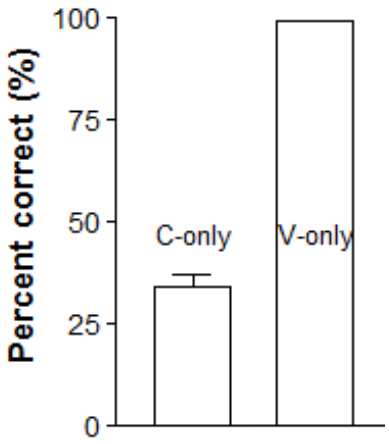
1.2.2 Segmental impacts: Vowels vs. consonants

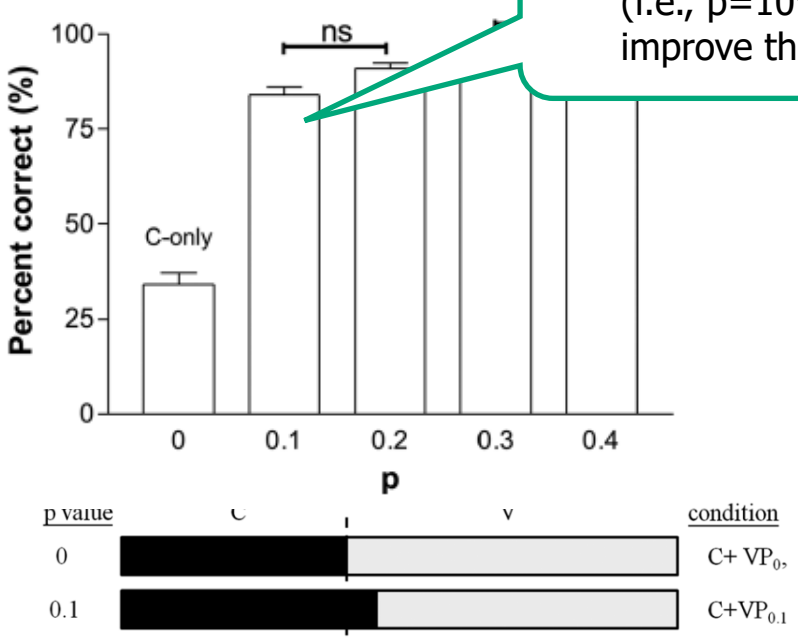
- Noise-replacement paradigm
- A remarkable (2:1) advantage of vowels (Vs) vs. consonants (Cs) for English sentence perception (Cole et al., 1996; Kewley-Port et al., 2007).
  - 87.4% (V-only sentences, with Cs replaced by white noise)
  - 46.6% (C-only sentences, with Vs replaced by white noise)



Segmental impacts: Mandarin speech perception

a ratio of 3



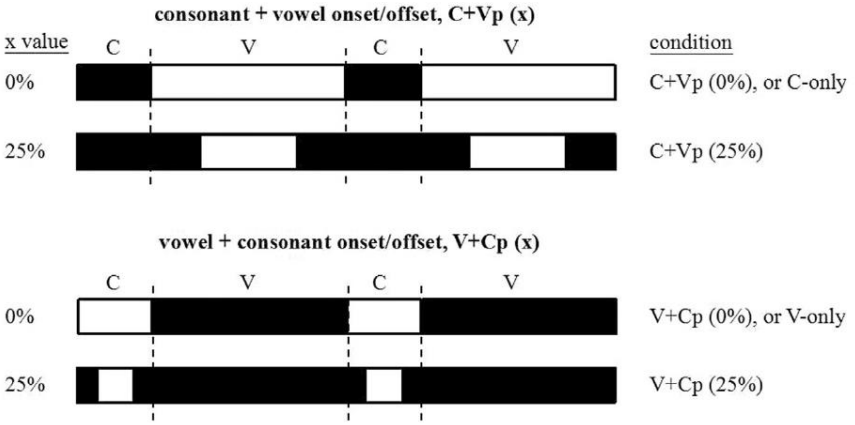


1. A small amount of vowel edge (i.e., p=10%) can significantly improve the perception.

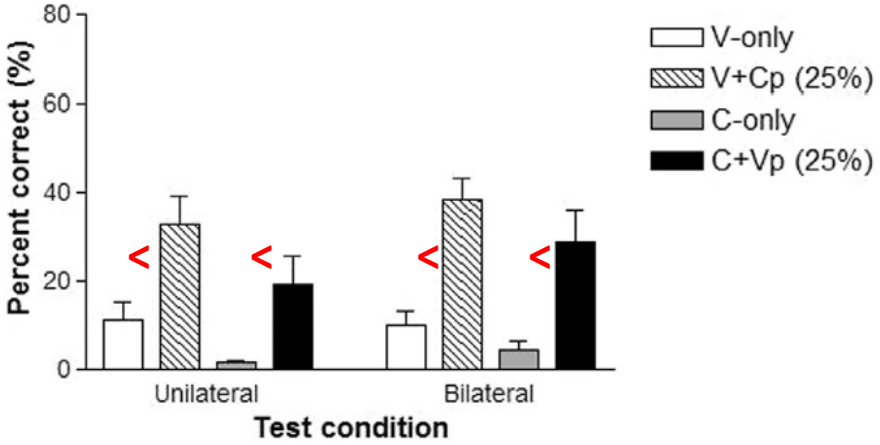
- Implications:
  - Vowels carry more intelligibility importance in Mandarin than in English.
  - Vowel-consonant transition carries important information for Mandarin speech perception.

"Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility," JASA, 2013

Segmental impacts: CI speech perception



Participants: 11 CI users

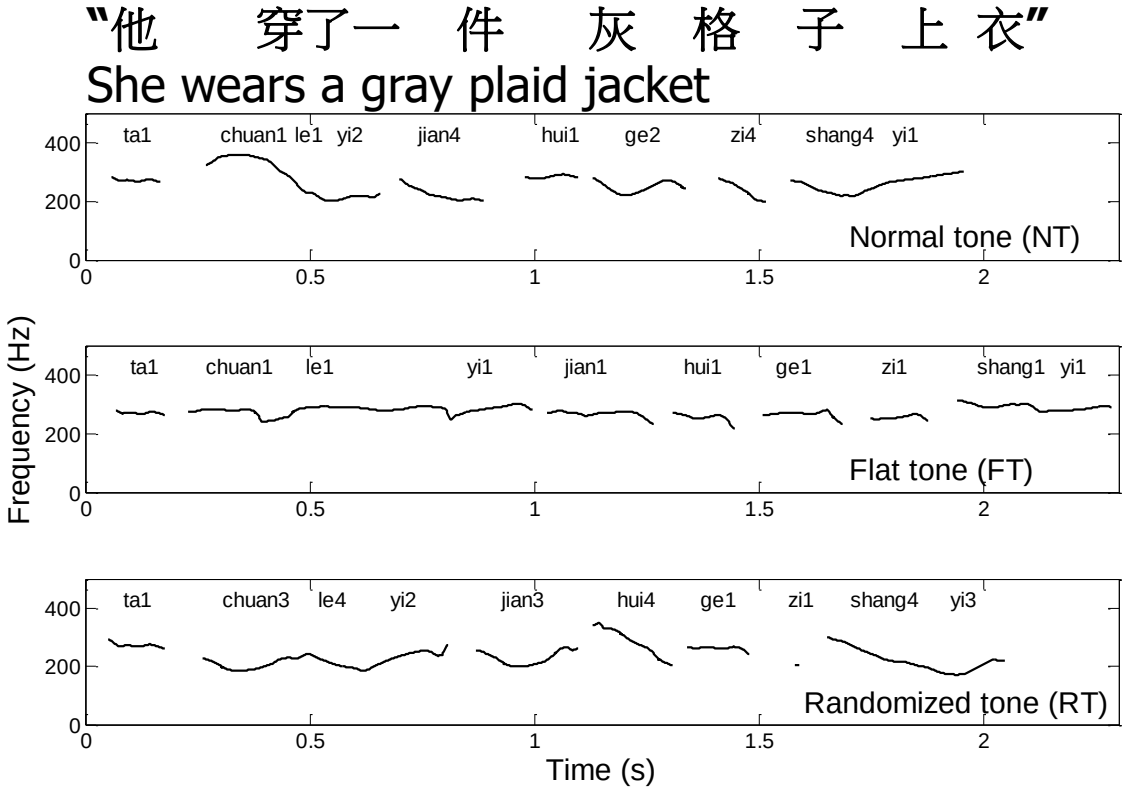


- Implications:
  - Perceptual advantage of vowels over consonants under both unilateral and bilateral CI speech perception.
  - Adding V-C transitions can significantly improve speech perception.

"Segmental contribution to cochlear implant speech perception," Speech Communication, 2019

1.2.3 F0 contour for Mandarin speech perception

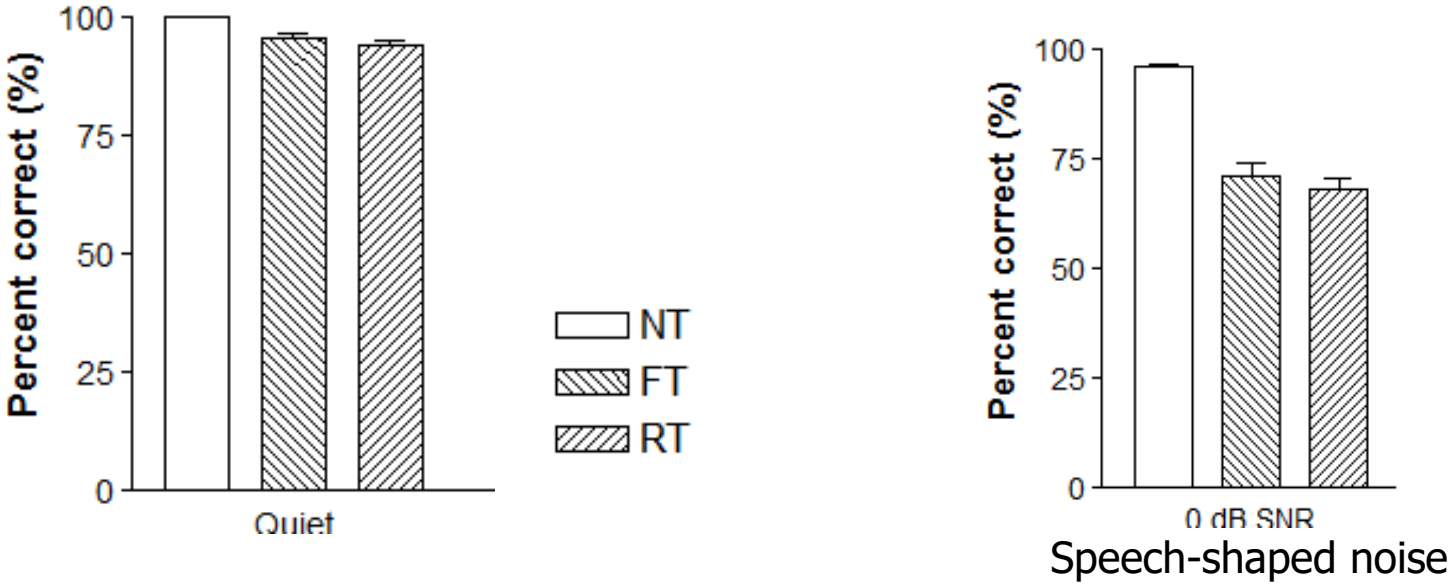
- Experiment design with a text-to-speech (TTS) engine
  1. Normal tone (NT)
  2. Flat tone (FT)
  3. Randomized tone (RT)



"Effects of lexical tone contour on Mandarin sentence intelligibility," Journal of Speech, Language, and Hearing Research, 2014

F0 contour for Mandarin speech perception

1) Normal-hearing adults



- Implications:
  - For assistive hearing devices, more efforts need to be focused on how to effectively deliver the tonal information in challenging listening conditions, e.g., in noise.

"Effects of lexical tone contour on Mandarin sentence intelligibility," Journal of Speech, Language, and Hearing Research, 2014

F0 contour for Mandarin speech perception

2) Children with cochlear implants

Participants

- CI group: 33 children (18 boys and 15 girls), aged between 3 and 6.
- Normal hearing group: 33 (17 boys and 16 girls) age-matched
- All participants were of normal intelligence

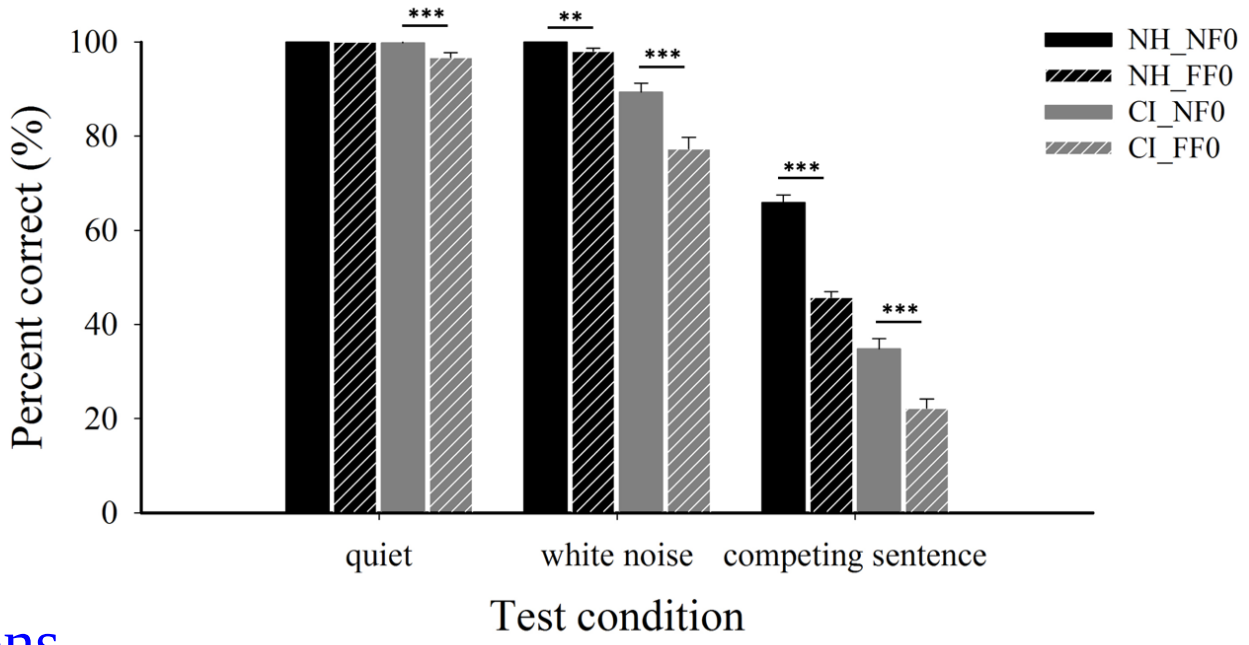
Stimuli

- Mandarin pediatric sentence recognition (MPSI) test (Zheng et al., 2009)
  - 2 testing conditions: sentences with natural F0 contours (NF0) or sentences with flat F0 contours (FF0).
- Two types of noise, i.e., white noise and competing sentence.

"Effects of fundamental frequency contours on sentence recognition in Mandarin-speaking children with cochlear implants," Journal of Speech, Language, and Hearing Research, 2020

F0 contour for Mandarin speech perception

2) Children with cochlear implants



Implications

1. Despite the limitation on F0 extraction in CI systems, F0 contours still play a major role in sentence recognition in both quiet and noise among pediatric implantees.
2. The contribution of F0 contour is even more salient than that in age-matched NH children.

"Effects of fundamental frequency contours on sentence recognition in Mandarin-speaking children with cochlear implants," Journal of Speech, Language, and Hearing Research, 2020

Outline

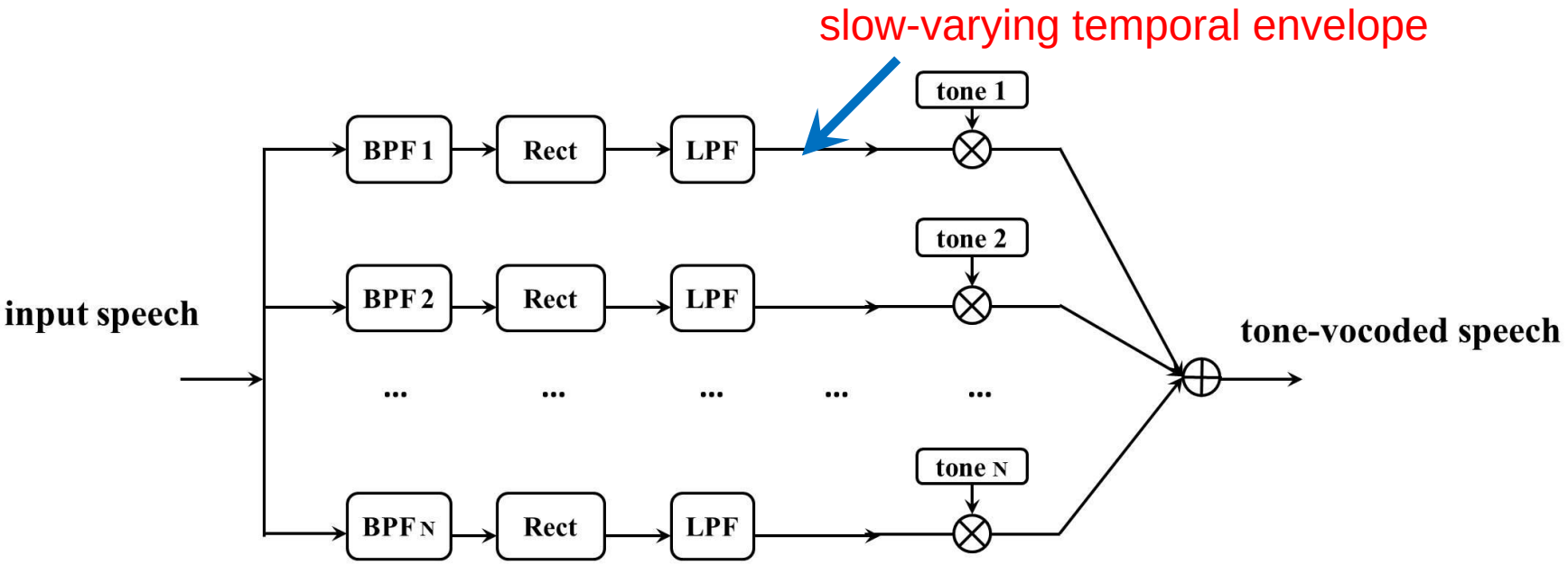
1. Background
   - Cochlear implants (CIs)
   - Acoustic cues for speech perception
2. Cochlear implants speech perception
   - Vocoder model for simulating CI speech perception
   - Combined acoustic-electric stimulation
   - Objective intelligibility evaluation for CI speech perception
3. Summary

2.1 Vocoder model for simulating CI speech perception

- Vocoder model uses multichannel envelope waveforms to synthesize envelope-based speech.
- Vocoder-based simulations have been used widely as an effective tool for assessing the effects affecting CI speech perception in the absence of patient-specific confounds.
  - effects of number of CI channels, envelope cutoff frequency, F0 discrimination, electrode insertion depth, filter spacing, background noise, etc.
  - predict well the pattern or trend in performance in CI users.
  - not to predict the absolute levels of performance of individual users, but rather the trend in performance when a particular speech-coding parameter is varied.

Tone-vocoder model for envelope-based speech synthesis

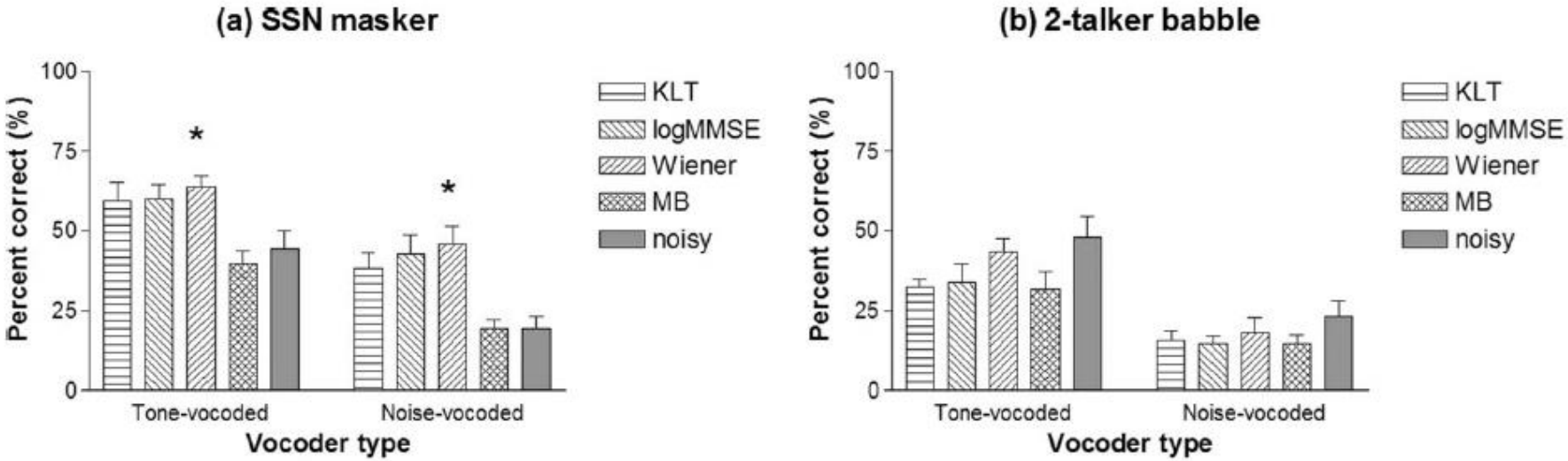
- Tone vocoder and tone-vocoded speech



BPF: band-pass filtering LPF: low-pass filtering
Rect: Wave rectification

"Effects of noise suppression and envelope dynamic range compression to the intelligibility of vocoded sentences for a tonal language," JASA, 2017

CI simulation: 1) Effect of carrier signal

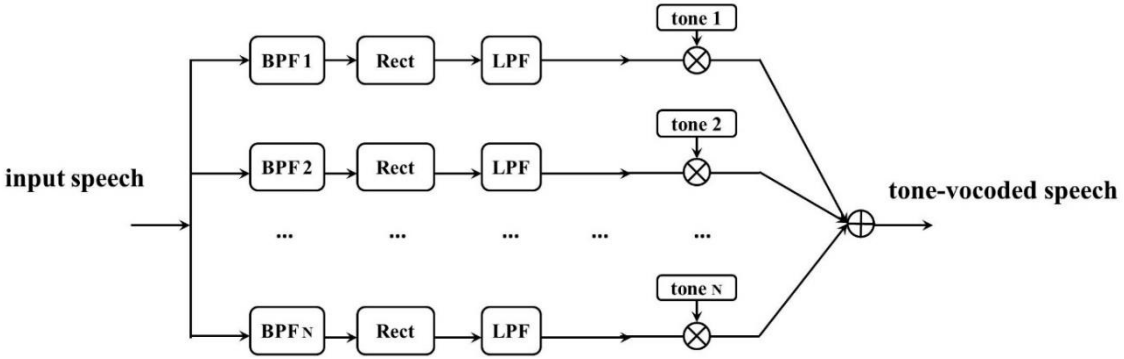


Implications

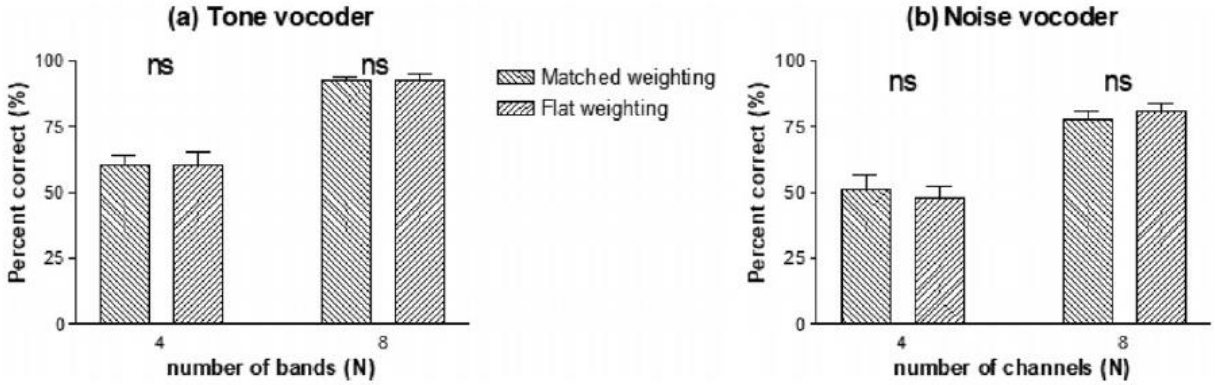
1. Tone-vocoded speech is more intelligible than noise-vocoded speech.
   - The spectral sidebands contained in TV speech when a pure tone is multiplied with the envelope waveform.
   - White-noise carriers have intrinsic envelope fluctuations that are absent in pure tone carriers

"Effects of noise suppression and envelope dynamic range compression to the intelligibility of vocoded sentences for a tonal language," JASA, 2017

CI simulation: 2) Effect of channel weighting



- Flat weighting: equal power weight across all channels
- Matched weighting: according to the power of each channel



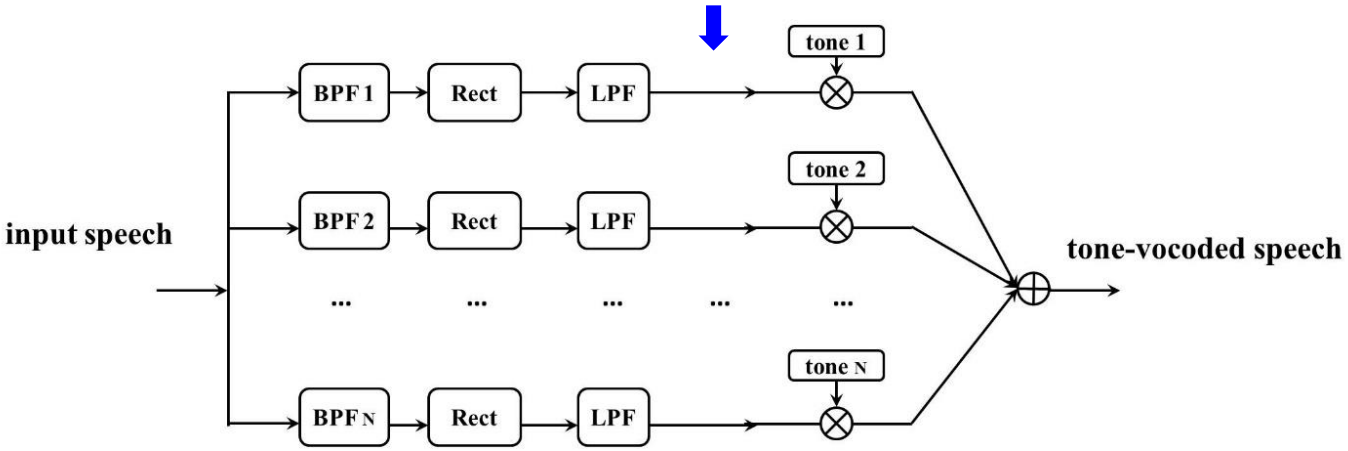
ns: non-significant

Implication

- Band power weighting did not significantly affect the intelligibility of stimuli synthesized with temporal information from a few bands.

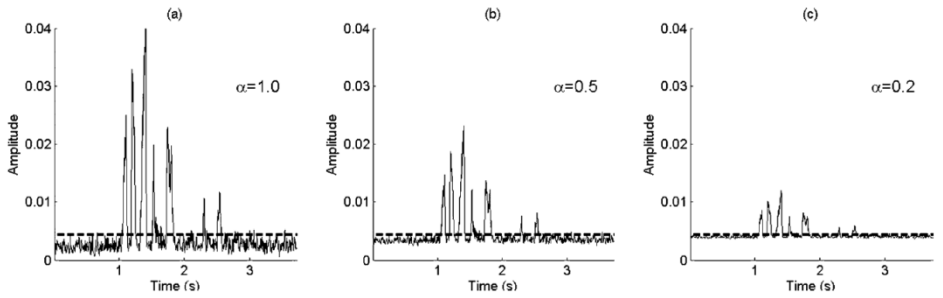
"Effect of band power weighting on understanding sentences synthesized with temporal information," JASA, 2019

CI simulation: 3) Effect of envelope dynamic range



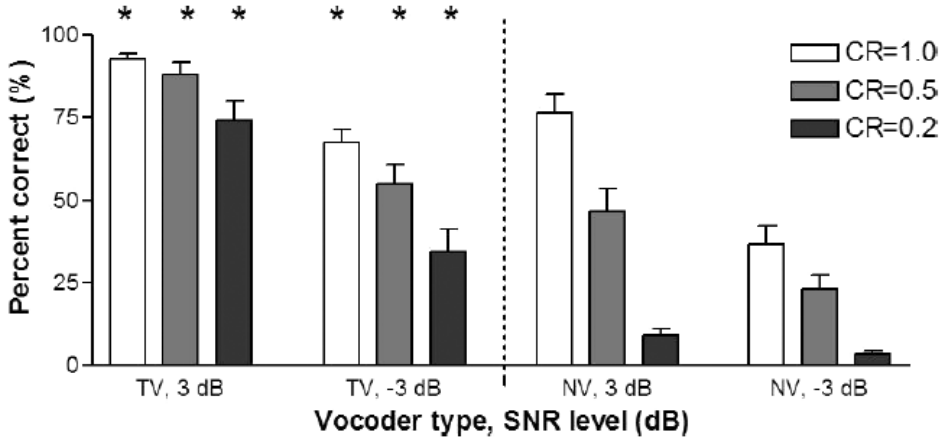
$$y = \alpha \times (x - \bar{x}) + \bar{x},$$

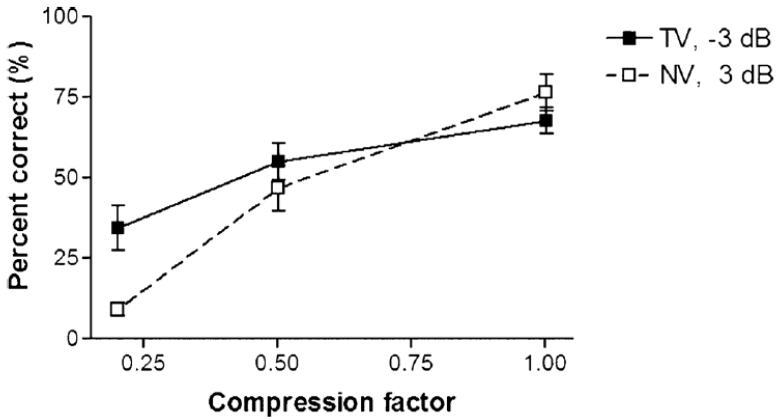
- When a=0, the compressed amplitude envelope becomes a direct current (dc) signal with a constant value of x, and the dynamic range is 0 dB.
- When a=1.0, the output amplitude envelope maintains the original dynamic range of the input (i.e., no envelope compression).



"Effects of noise suppression and envelope dynamic range compression to the intelligibility of vocoded sentences for a tonal language," JASA, 2017

CI simulation: 3) Effect of envelope dynamic range





Implications

- While the envelope dynamic range was narrowed, both TV and NV speech showed reduced intelligibility performance.
- However, NV speech was more negatively affected by envelope dynamic range compression, yielding a substantial intelligibility gap between TV and NV speech.

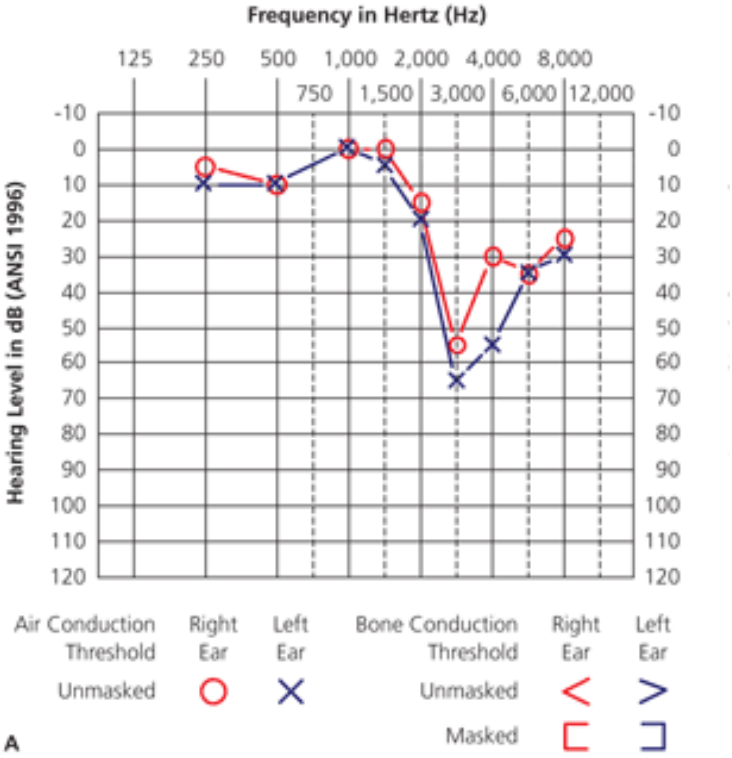
"Effects of noise suppression and envelope dynamic range compression to the intelligibility of vocoded sentences for a tonal language," JASA, 2017

Outline

1. Background
   - Cochlear implants (CIs)
   - Acoustic cues for speech perception
2. Cochlear implants speech perception
   - Vocoder model for simulating CI speech perception
   - Combined acoustic-electric stimulation (EAS)
   - Objective intelligibility evaluation for CI speech perception
3. Summary

Low-frequency residual hearing

For persons with high levels of low-frequency residual hearing, a further relaxation in the criteria for implantation is to use combined electric-acoustic stimulation (EAS).



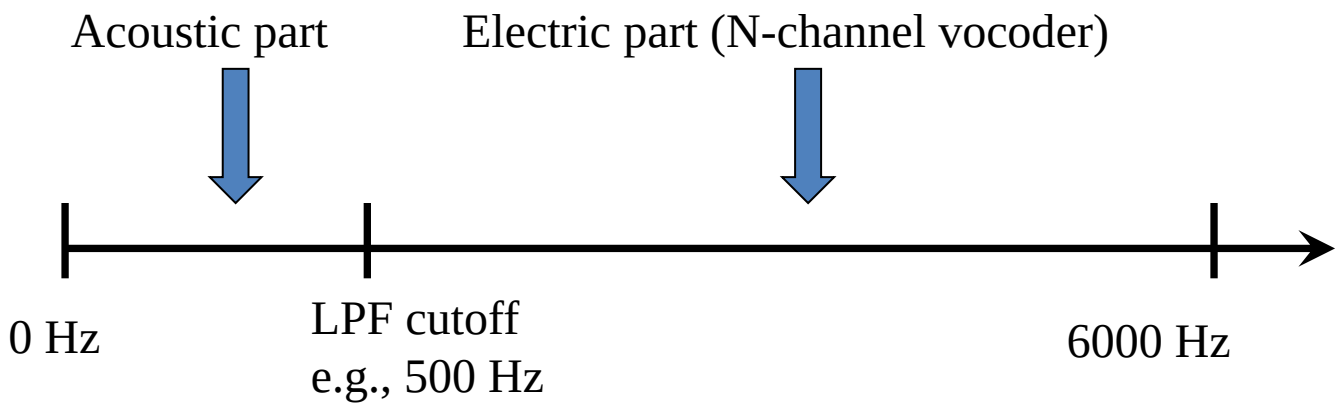
Combined electric-acoustic stimulation (EAS)

- In EAS patients, an electrode array is implanted only partially into the base region of cochlea so as to preserve the residual acoustic hearing at low frequencies, which many patients still have.
- The low-frequency and high-frequency (e.g., >1000 Hz) speech information is provided to these patients via a hearing aid and a CI, respectively.

http://en.wikipedia.org/

Combined-stimulation advantage

- A substantial of evidences are supporting the benefits of EAS in terms of better speech recognition, especially in noisy environment.
  - The combined-stimulation advantage refers to an improvement in speech recognition when electric stimulation in CI is supplemented by LF acoustic information, i.e. CI+LF vs. CI-only.
- Simulating EAS signal processing



EAS: 1) Effect of F0 contour

Motivation

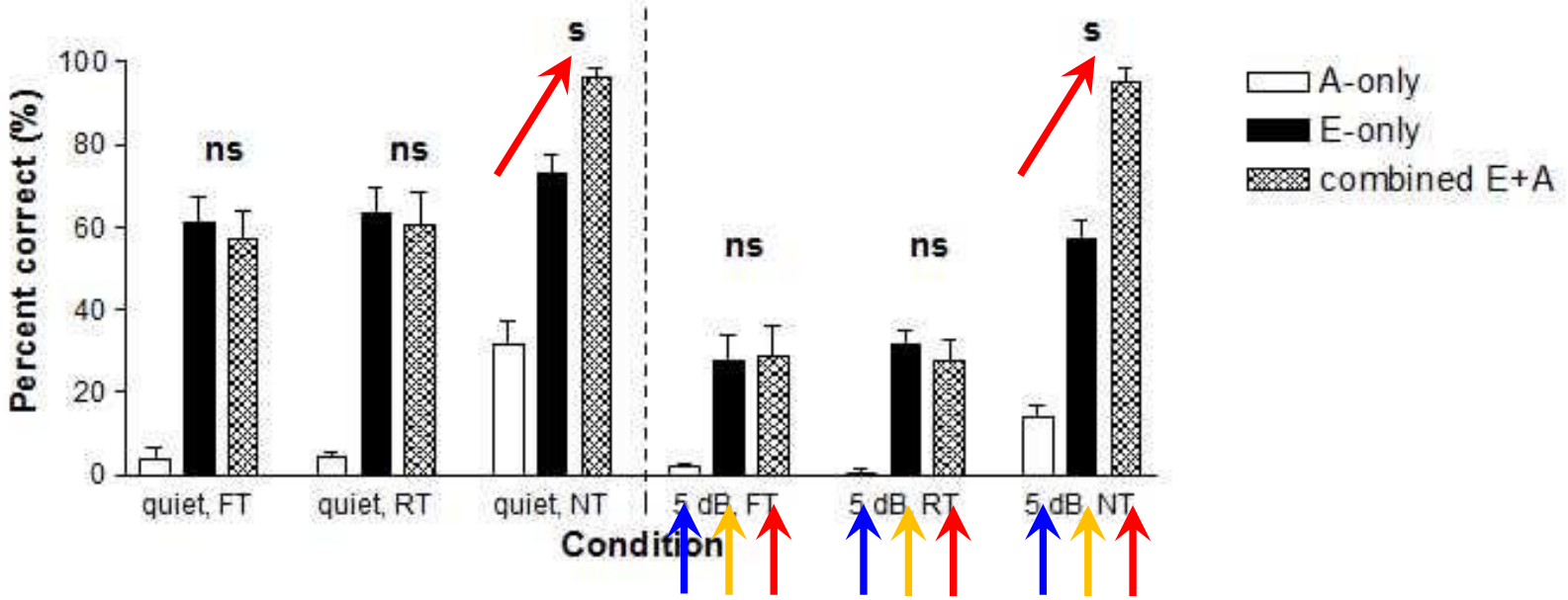
- To understand the effect of F0 contour on understanding Mandarin sentences in EAS hearing.

Methods

- First manipulate F0 contour in three conditions: NT (normal tone), FT (flat tone), and RT (randomized tone).
- Then synthesize three test conditions
    - A-only: simulating low-frequency residual hearing
    - E-only: simulating CI speech processing
    - combined E+A: simulating EAS hearing

"Effects of fundamental frequency contour on understanding Mandarin sentences in bimodal hearing simulations," JASA, 2018

EAS: 1) Effect of F0 contour



Implications

1. Correct tonal information is important for understanding **A-only** speech.
2. Correct tonal information is important for understanding **E-only** (vocoded) speech, simulating CI.
3. Correct tonal information is important for understanding **combined E+A** speech, simulating EAS hearing.
4. EAS benefit is seen only with correct tonal information.

"Effects of fundamental frequency contour on understanding Mandarin sentences in bimodal hearing simulations," JASA, 2018

EAS: 2) Effect of speech segments

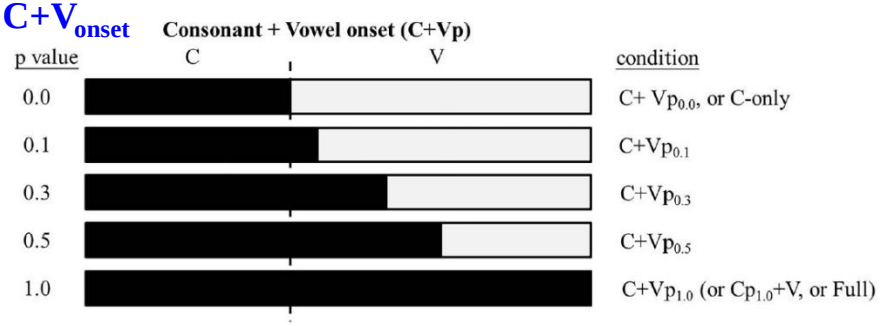
Motivation

- To examine how segmental cues (e.g., vowels, vowel-consonant transitions) interact with the E+A advantage.

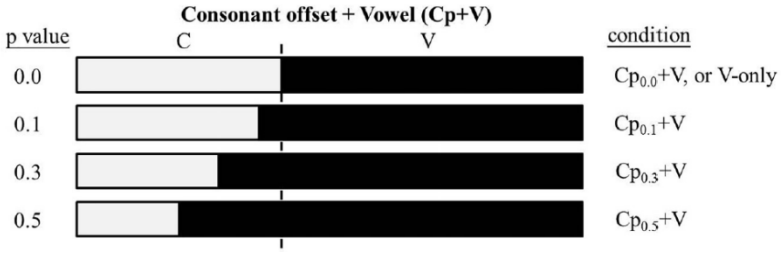
Methods

- First generate noise-replaced stimuli, selectively preserve target speech segments, and replace the rest segments with white noise.

(A) $C+V_{onset}$



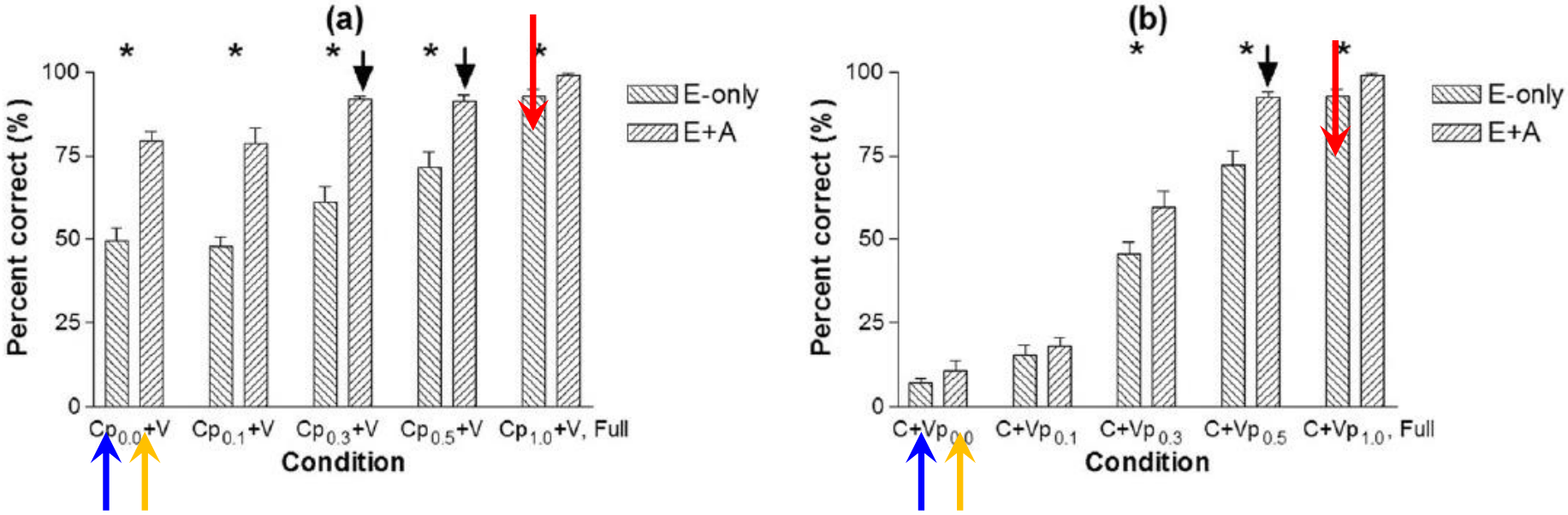
(B) $C_{offset}+V$



- Then synthesize two test conditions
  - E-only
  - combined E+A

"Perceptual contributions of vowels and consonant-vowel transitions in simulated electric-acoustic hearing," JASA, 2019

EAS: 2) Effect of speech segments



Implications

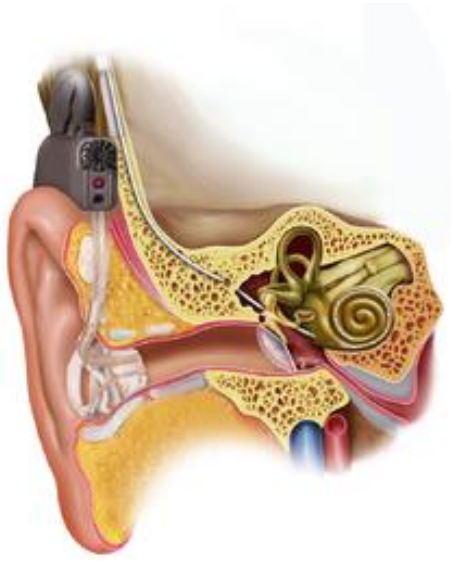
1. V-only sentences are more intelligible than C-only sentences under both E-only and E+A conditions.
2. The addition of a portion of C-V transitions may significantly improve the intelligibility of V-only and C-only sentences in combined EAS, yielding sentence understanding performance equivalent to that of **E-only condition with all speech segments**.

"Perceptual contributions of vowels and consonant-vowel transitions in simulated electric-acoustic hearing," JASA, 2019

EAS: 3) Effect of temporal misalignment

- Could be caused by at least two reasons, i.e., processing delay difference and inherent biological delay of CI and HA speech processing.

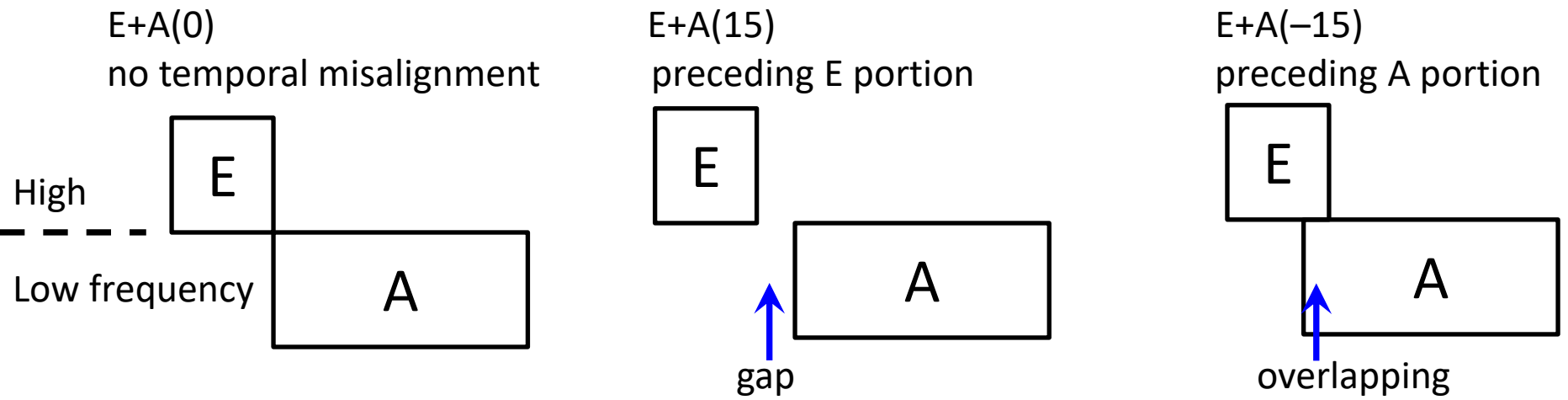
  1. Separate CIs and HAs may have different processing delays and are not designed specifically to work together, which yields a difference of processing delay between a CI and a HA (e.g., Francart and McDermott, 2013).
  2. The electrical stimulation in CI occurs almost instantly, while the acoustic sound from HA is presented through the ear canal and middle ear, and takes time to travel the basilar membrane to the apical portion of the cochlea.



http://en.wikipedia.org/

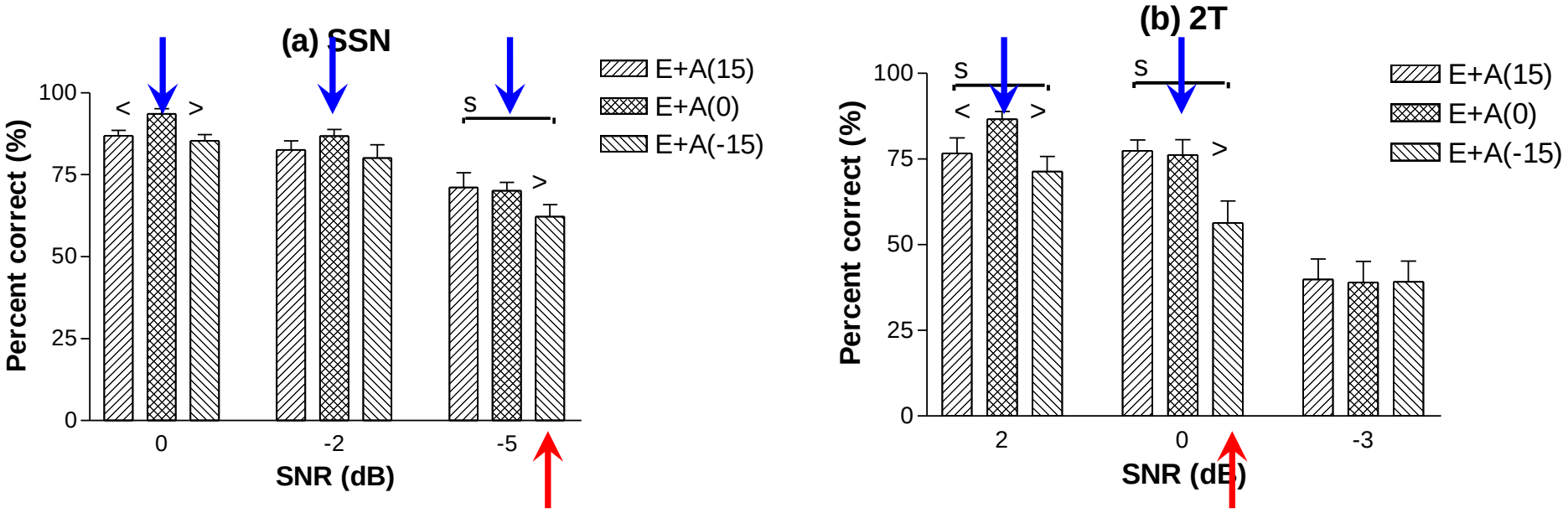
EAS: 3) Effect of temporal misalignment

- Methods
  - a relative time shift was added between the E and A portions in EAS, including 15, 0, and –15 ms, yielding three E+A conditions of E+A(15), E+A(0), and E+A(–15).



"Effect of temporal misalignment on understanding Mandarin sentences in simulated combined electric-and-acoustic stimulation," JASA, 2020

EAS: 3) Effect of temporal misalignment

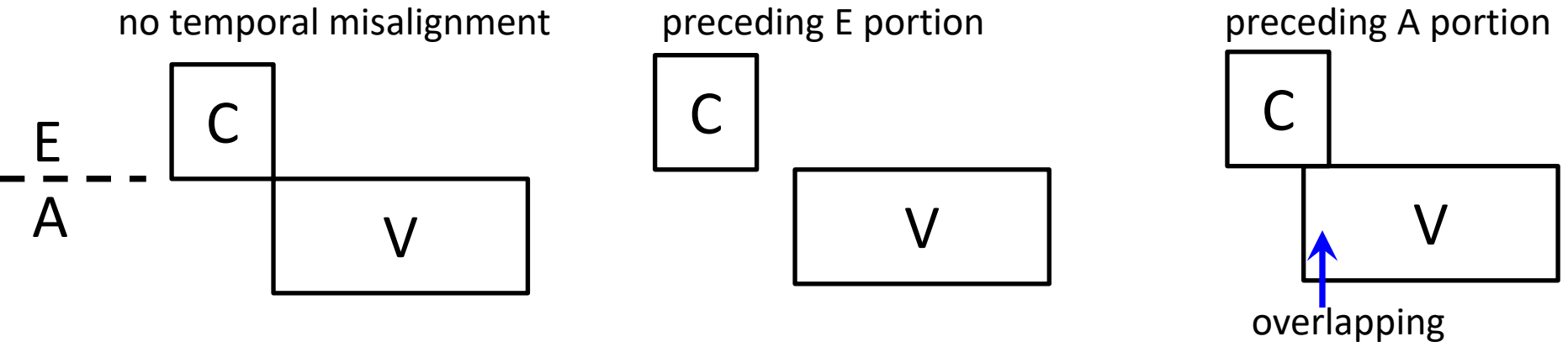


- Temporal misalignment may potentially cause a negative influence on the sentence perception task.
- The relative temporal misalignment between the E and A portions may have a different influence on the combined EAS advantage.

"Effect of temporal misalignment on understanding Mandarin sentences in simulated combined electric-and-acoustic stimulation," JASA, 2020

EAS: 3) Effect of temporal misalignment

- A relatively better understanding under the E+A condition with the preceding E portion (or a worse performance under the E+A condition with the preceding A portion).
  - Mandarin words have a mono-syllabic structure, C+V, where C is normally at high-frequency (E portion in EAS) and V is at low-frequency (A portion in EAS).
  - C-V transitions carry important information for speech perception, particularly in challenging conditions.



"Effect of temporal misalignment on understanding Mandarin sentences in simulated combined electric-and-acoustic stimulation," JASA, 2020

Outline

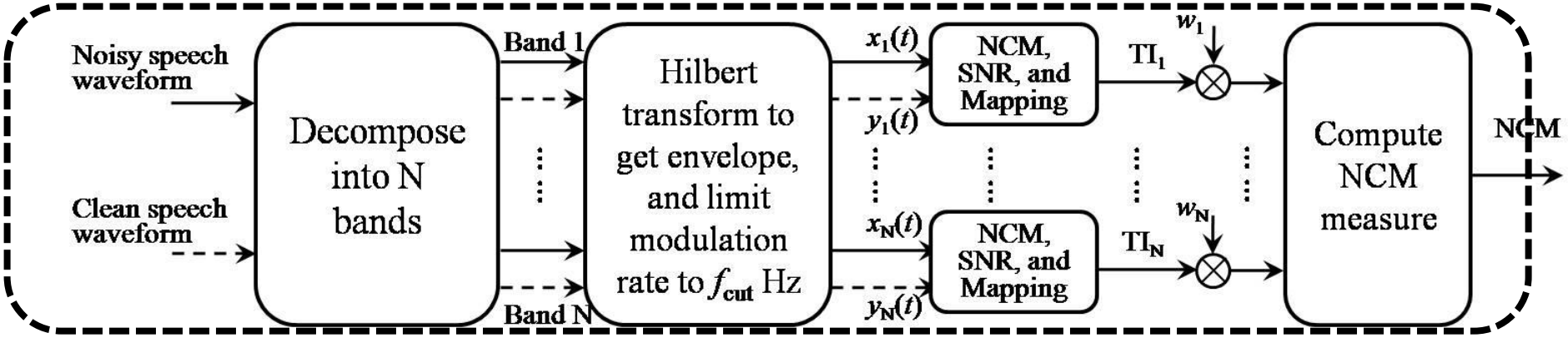
1. Background
   - Cochlear implants (CIs)
   - Important cues related to CI speech perception
2. Cochlear implants speech perception
   - Vocoder model for simulating CI speech perception
   - Combined acoustic-electric stimulation
   - Objective intelligibility evaluation for CI speech perception
3. Summary

Intelligibility evaluation for CI speech perception

- Two categories: subjective evaluation and objective evaluation.
- Subjective 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 intelligibility evaluation aims to provide a numerical index on speech intelligibility via a computational model.

2.3.1. Envelope-based intelligibility index

- Normalized-covariance measure (NCM)



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

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

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

$$\mathrm{TI}_i = \frac{\mathrm{SNR}_i + 15}{30}$$

(TI: transmission index, and ≤ 1)

Goldsworthy et al., JASA, 2004

Predicting the intelligibility of vocoded speech

Table 1. Summary of subject and test conditions involved in the correlation analysis

Experiment	No. Subjects	Maskers	SNR Levels (dB)	No. Conditions: Tone Vocoder	No. Conditions: EAS Vocoder
1	7	SSN, two-talker	5, 0, −5	6	24
2	6	SSN, two-talker	5, 0	4	20
3	9	SSN	5, 0	6	20

a total of 80 tone- and EAS-vocoding conditions

Table 2. Filter cutoff (−3 dB) frequencies used for the tone- and EAS-vocoding processing

Channel	Tone Vocoding: Low (Hz)	Tone Vocoding: High (Hz)	EAS Vocoding: Low (Hz)	EAS Vocoding: High (Hz)
1	80	221		
2	221	426	Unprocessed (80–600)	
3	426	724		
4	724	1158	724	1158
5	1158	1790	1158	1790
6	1790	2710	1790	2710
7	2710	4050	2710	4050
8	4050	6000	4050	6000

Table 4. Correlation coefficients obtained with the NCM measure for different modulation rates ranging from 12.5 to 180 Hz

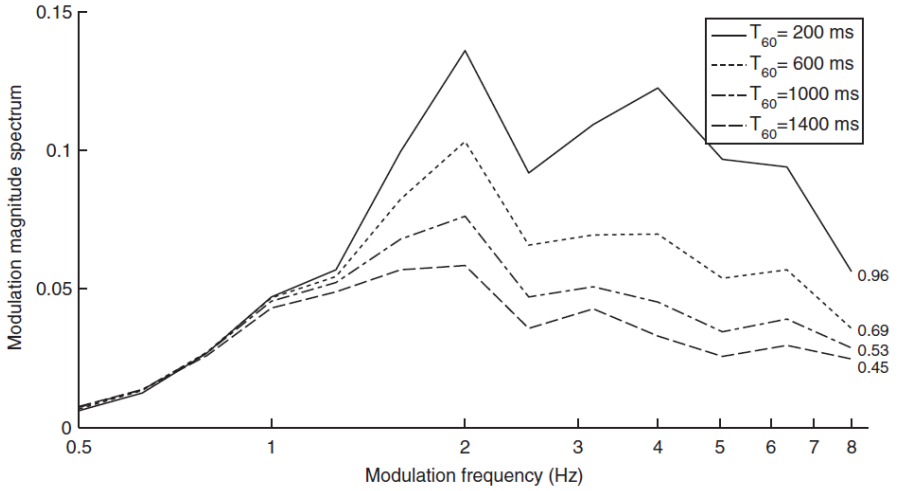
Modulation Rate (Hz)	*r*
12.5	0.85
20	0.90
31.5	0.90
40	0.90
60	0.91
80	0.91
100	0.92
140	0.92
180	0.92

The Pearson's correlation coefficient (*r*) between subjective scores and objective measures was used to assess the prediction performance.

"Predicting the intelligibility of vocoded speech," Ear and Hearing, 2011

2.3.2 Nonintrusive speech intelligibility measure: the Average Modulation-spectrum Area (ModA)

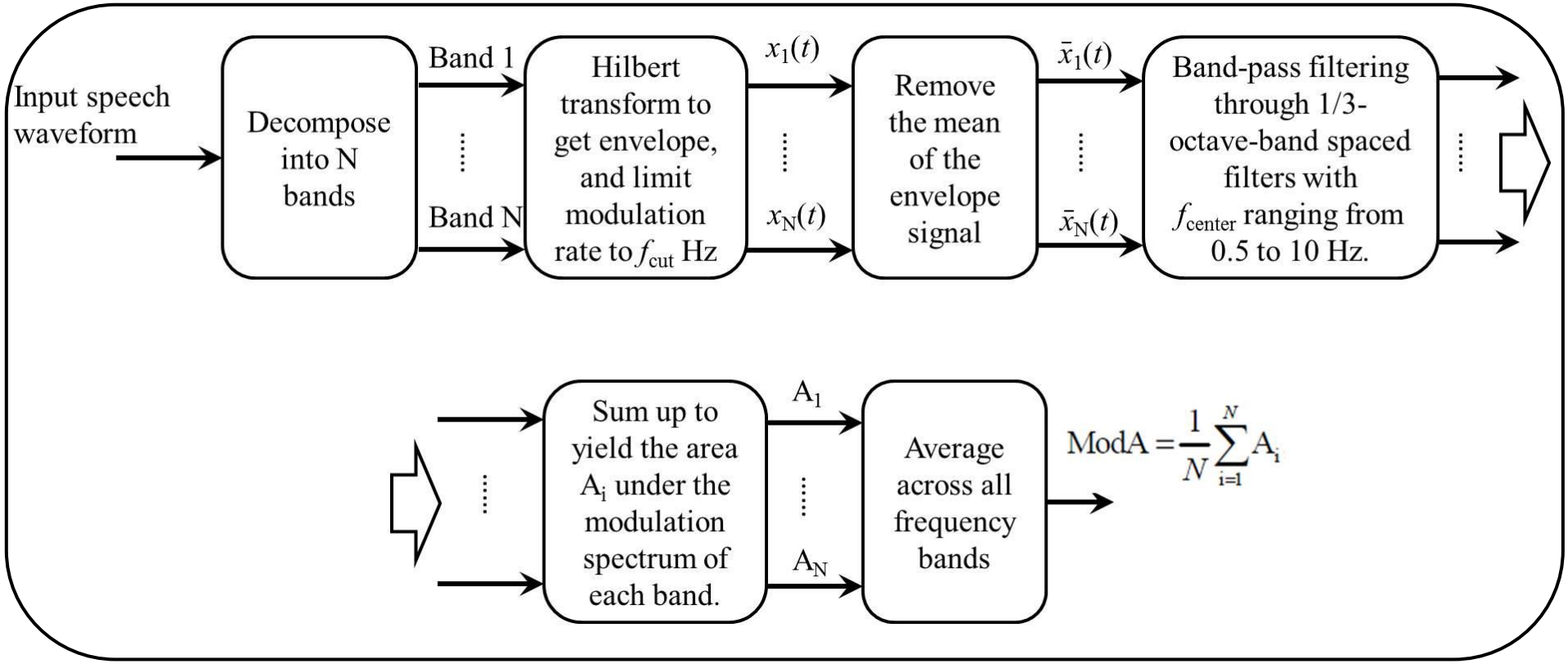
- Most of the previous indices are intrusive. That is, a clean reference is needed.
  - In most scenarios, such reference signal is not available.
  - Non-intrusive (or reference-free) index only uses degraded speech signal to predict intelligibility.
- As the level of reverberation is increased, the modulation spectrum of the reverberant envelopes becomes flat and shifts down.



Value of area

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

"Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," Biomedical Signal Processing and Control, 2013


Input speech waveform
Decompose into N bands
Band 1
Band N
Hilbert transform to get envelope, and limit modulation rate to f_{cut} Hz
$x_1(t)$
$x_N(t)$
Remove the mean of the envelope signal
$\bar{x}_1(t)$
$\bar{x}_N(t)$
Band-pass filtering through 1/3-octave-band spaced filters with f_{center} ranging from 0.5 to 10 Hz.
Sum up to yield the area A_i under the modulation spectrum of each band.
A_1
A_N
Average across all frequency bands
$\text{ModA} = \frac{1}{N}\sum_{i=1}^{N} A_i$

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)

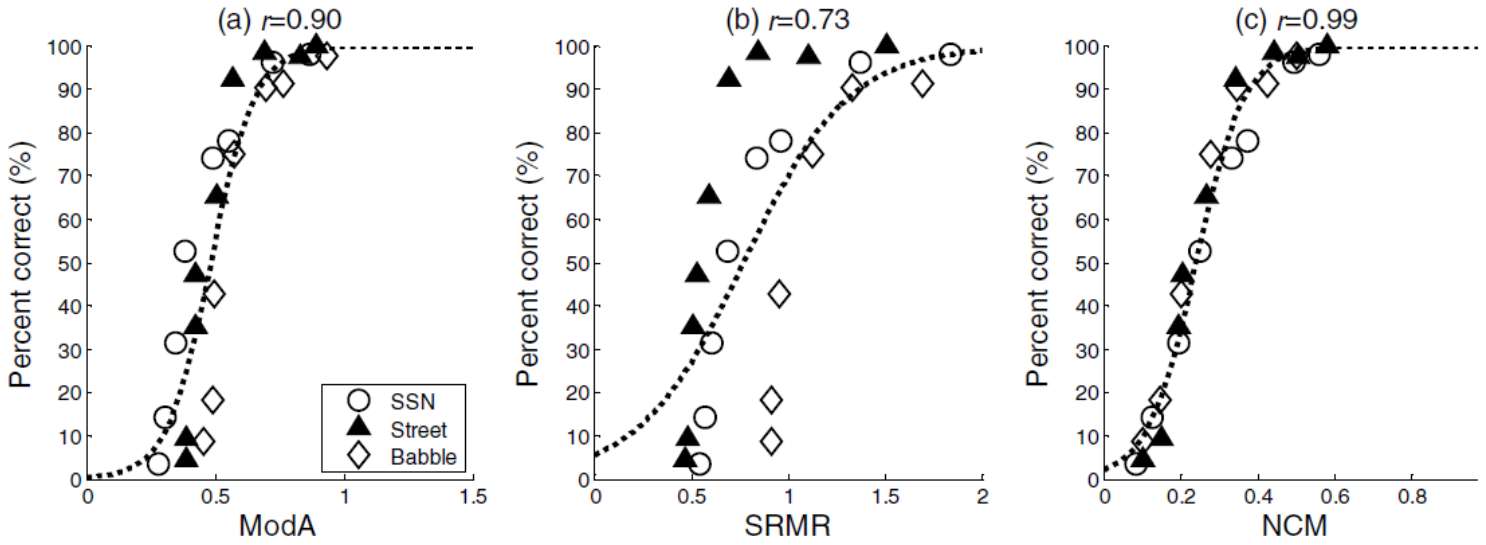


Fig. 2. Scatter plots of sentence recognition scores against the (a) ModA, (b) SRMR, and (c) NCM values.

"Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," Biomedical Signal Processing and Control, 2013

Outline

1. Background
   - Cochlear implants (CIs)
   - Important cues related to CI speech perception
2. Cochlear implants speech perception
   - Vocoder model for simulating CI speech perception
   - Combined acoustic-electric stimulation
   - Objective intelligibility evaluation for CI speech perception
3. Summary

Summary

1. A lot of works are needed to further understand the mechanism of CI speech perception.
   - Effects of acoustic and linguistic cues
2. Vocoder model is a valuable tool to study CI speech perception.
   - Simulating CI and EAS speech perception
3. Objective intelligibility evaluation could help us design novel speech processing and coding approaches for improving CI speech perception.