THE ROLE OF TEMPORAL CUES IN SPEECH PERCEPTION: A SYSTEMATIC REVIEW

Ms. Saranya Arya Mundayoor

Registration No. 19AUD033

A Dissertation Submitted in Part Fulfilment of Degree of Master of Science

(Audiology)

University of Mysuru

Mysuru

ALL INDIA INSTITUTE OF SPEECH AND HEARING
MANASAGANGOTHRI, MYSURU-570006
September 2021

CERTIFICATE

This is to certify that this dissertation titled “**The Role of Temporal Cues in Speech Perception: A Systematic Review**” is a bonafide work submitted in part fulfilment for the degree of Master of Science (Audiology) by the student holding Registration Number: 19AUD033. This has been carried out under the guidance of a faculty member of this institute and has not been submitted earlier to any other University for the award of any other Diploma or Degree.

Mysuru.

September, 2021

Dr. M Pushpavathi

Director

All India Institute of Speech and Hearing
Manasagangothri, Mysuru-570006

CERTIFICATE

This is to certify that this dissertation titled "**The Role of Temporal Cues in Speech Perception: A Systematic Review**" is a bonafide work submitted in part fulfilment for the degree of Master of Science (Audiology) by the student holding Registration Number: 19AUD033. This has been carried out under my supervision and guidance. It is also certified that this dissertation has not been submitted earlier to any other University for the award of any other Diploma or Degree.

Mysuru.

September, 2021

Dr. Ajith Kumar U.

Guide

Professor,
Department of Audiology
All India Institute of Speech and Hearing
Manasagangothri, Mysuru-570006

DECLARATION

This is to certify that this dissertation entitled “**The Role of Temporal Cues in Speech Perception: A Systematic Review**” is the result of my own study under the guidance of Dr. Ajith Kumar U., Professor, Department of Audiology, All India Institute of Speech and Hearing, Mysuru, and has not been submitted earlier to any other University for the award of any other Diploma or Degree.

Mysuru. Registration Number: 19AUD033

September, 2021

Dedicated to my dearest parents

ACKNOWLEDGEMENT

First and foremost, I would like to convey my sincerest gratitude to my professor and guide, **Dr. Ajith Kumar U.** – to be able to do my dissertation under the guidance of such a legend in the field of Audiology is nothing short of an honour and a privilege. One cannot ask for a better guide. Thank you so much for this invaluable opportunity, and for all the patience and support, sir!

Words are not enough to express my gratitude to my loving parents, **Savithri** and **Sathish Mundayoor**, and my dearest brother, **Amol Mundayoor**. Without the unfailing support, encouragement and love that you have always given me, none of this would have been possible. Thank you for all that you've done for me.

My dearest friend **Shinsi**, you are an absolute gem. You always know exactly what to say and your jokes never fail to make me smile, and I know I can always count on you. Thank you for always being there for me.

Chethan, you truly are a most wonderful dissertation partner and friend. Thank you for all the help throughout this journey, macha!

My girls, **Ashitha** and "**Sunnimeeryoles**" (**Hasla**, **Anshaba**, **Joel**, **Shinsi**, **Shaime** & **Ashu**) – y'all are sunshine. Whether it's just having tea in the mess or on a vacation somewhere far away, there's never a dull moment with you peeps. You guys have my heart <3

My **"Crazy Girls"** (**Sneha**, **Aadi**, **Malu**, **Lechu**, **Abna**, **Aishwarya**, **Shinsi** & **Dyuthi**), you guys are love. This place is a hundred times better because of you all.

This list would be incomplete without the mention of you peeps - **Prabuddha**, you are the champagne of friends, and I know you get this ref; **Kruthika** and **Pretek**, y'all are two of the purest things I’ve ever come across (Special thanks to you three for patiently putting up with all of my BS); **Kavitha**, **Shejal**, **Anju**, **Tasneem**, **Kajol** and **Anima**, you’re the kind of people Olaf was talking about.

Abhishek, **Aashish**, **Ankit, Sunny, Zohra, Namitha** & **Vidya,** thank you for making postings, classes and non-postings/classes so much more enjoyable.

A huuuge thanks to all my **Section B** buddies – I’ve had the best time with you peeps, thank you so much!

Last but not least, **Renovators 2.0** and **Renovators**, thank you for all the memories!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Traditionally, speech perception was thought to depend primarily on the spectral information represented in the speech signal. The apparent success demonstrated in speech understanding by wearers of single-channel cochlear implants (Tyler, 1988) was one of several reasons that brought to the notice of the researchers the importance of temporal information in speech. Single-channel cochlear implants rely on the electrical stimulation provided to a single electrode at a certain region in the cochlea such as the promontory; therefore, this system cannot provide frequency-specific information to the auditory system. In the absence of spectral information, temporal cues must be responsible for speech perception.

There are considerable psychophysical and physiological evidences to show that temporal cues are extracted in the cochlea. The basilar membrane of the normal cochlea, with its fine frequency resolution, can be picturized as being composed of an array of bandpass filters arranged tonotopically along its length. At the output of these bandpass filters, the broadband speech stimulus is divided into a series of narrowband signals. Each of these signals are further separated into its temporal envelope (E), which consists of slowly varying amplitude information (the AM component), and its temporal fine structure (TFS), the rapidly varying frequency information with rate close to the centre frequency of the band (the FM component) (Moore, 2008; Rosen, 1992). The E is therefore said to modulate the carrier/TFS.

More formal definitions have been put forth for temporal cues in speech. Rosen (1992), for example, proposed three categories of temporal cues based on the frequency range over which they exist: the envelope cue (2-50Hz), the periodicity cue (50-500Hz) and the temporal fine structure cue (600-10kHz). According to Rosen, the envelope cue conveys information on manner, tempo, rhythm and syllabicity; the periodicity cue is important in delineating voicing, stress and intonation; the fine structure cue, place and vowel quality information. Several other linguistic contrasts may also be conveyed by each of these cues, but their roles are assumed to be weak (Rosen, 1992). Apart from Rosen's definition, others exist and have been used in studies of temporal cues in speech perception. Reviewing them is beyond the scope of the current study and the interested reader is referred to Hilbert (1912).

Research conducted towards the goal of identifying the exact roles played by temporal cues has used several kinds of stimuli and focussed on different categories of the temporal information outlined above. Techniques employed in these studies generally involve cochlear implant simulations with vocoded speech using noise, sinusoid and/or pulse carriers, and auditory chimeras.

Auditory chimera, a hybrid stimulus that contains the envelope of one signal and the fine structure of another, was developed by Smith and co-workers in 2002. In a speech-speech chimera, which used the envelope of one speech signal to modulate the fine structure of another, the listeners reported hearing the speech coded in the envelope. In contrast, when a melody-melody chimera was used, it was the melody in the fine structure that was perceived by the listeners (Smith et al., 2002). Tonal languages may be expected to follow the latter cue, and the same was evidenced by Xu and Pfingst (2003), who found that TFS cues were predominantly involved in the identification of lexical tone in Mandarin Chinese.

Vocoded speech is effectively a special case of the auditory chimera, differing mainly in the fact that the stimulus used as the carrier is a non-speech stimulus (a noise, a sinusoid, or a pulse). Noise vocoding was used in one of the earliest and well-cited evidence in favour of E cues, by Shannon et al. (1995). Shannon and colleagues investigated the speech recognition ability in normal hearing individuals with stimuli having almost purely temporal information. Envelopes were extracted from the acoustic signal at different bandwidths by low-pass filtering at 16, 50, 160 and 500Hz to create noise-vocoded speech stimuli across one to four spectral bands. All eight participants recruited for the study were tested on 16 medial consonants, 8 vowels and simple sentences. The results revealed that performance improved as the number of bands increased from one to four. With just four bands, there was >85% correct identification of speech across all three conditions. It was also observed that except when vowels were the stimuli, scores reduced with the low-pass filtering condition of 16Hz, but were comparable across the other three cut-offs. Additionally, manner and voicing cues seemed to require only two spectral bands for maximum performance, but place cues were better identified as the number of bands increased (Shannon et al., 1995). Other studies have obtained similar results (Fogerty, 2011). These studies provide strong evidence for the importance of E cues in speech comprehension.

The relative contribution of E and TFS cues to speech perception has been extensively studied and the literature indicates that the auditory system predominantly depends on E cues in quiet (Shannon et al., 1995). The TFS cue may, by itself, contribute to speech perception in quiet (Gilbert et al., 2007; Gilbert & Lorenzi, 2006), although the effect does not seem to be as robust as that associated with the E cues. TFS cues aid in speech understanding in presence of noise and fluctuating maskers (Nelson et al., 2003), which render E cues inaccessible. It also helps in "glimpsing" speech in the short windows of favourable SNR (Lorenzi et al., 2006). New evidence indicates that TFS cues may facilitate stream segregation by tracking the fine spectral changes across both signal and masker (Apoux et al., 2013); it may also provide temporal information distinct from that provided by E cues, brought about by the fine spectral changes with time (Teng et al., 2019). The periodicity cue, although not directly implicated in facilitating intelligibility, may still be important in the identification of voicing contrasts (Faulkner et al., 2000).

Although TFS cues are reported to be important in speech perception in noise, the exact mechanism of how TFS contributes temporal information is not well understood. It has been suggested that the envelope may be recovered (presumably at the cochlea) from the fine structure when the auditory filters are narrow, and the analysis bands are wide and less in number (2 or lesser) (Ghitza, 2001). This recovery is said to take place in the normal human cochlea, because the auditory system is able to convert the rapidly fluctuating frequency information into slowly changing amplitude envelopes (Gilbert & Lorenzi, 2006). Several studies have, however, reported results to the contrary, indicating that the recovery of E from TFS cues is not enough to explain the speech perception abilities witnessed using TFS cues alone when the recovery mechanism is controlled for (Sheft et al., 2008; Teng et al., 2019). Another issue, related to the technology used to extract the E and TFS cues, is that both of these features of the speech signal are said to be inseparable by design; modifying the TFS inadvertently affects the E cues, and vice versa (Ghitza, 2001). This issue further complicates the matter of independent cue extraction.

Although it has been established that greater number of analysis bands are required to prevent E reconstruction from TFS, far too great a number may also be problematic. Zeng et al. (2003) has said that results with studies using greater than 16 bands should be interpreted with caution, since such finely-tuned filters may be associated with ringing and other related artifacts.

Another issue raised by Apoux and Healy has to do with “single carriers”. They argue that when the speech and noise signals are vocoded, the resulting carrier/TFS is a mixture of both the signal and the masker. The E now modulates a merged and non-distinct carrier, and the auditory system is no longer provided information from two different streams. To get around this issue, the authors have thus proposed an alternative solution: the masker-mixed-speech signal is chimerized into E and TFS which differ only in relative signal-to-noise ratio. This helps to study the relative contribution of each cue without majorly disrupting the other (Apoux & Healy, 2013). Other authors such as (Drullman, 1995) have similarly used novel stimuli and processing schemes.

Another widely used approach to separate E and TFS is the Hilbert transform. One major problem here is the nonseparation of the periodicity cues in the Hilbert E and TFS. This may confound the findings of studies that have used this technique to compare the relative contribution of envelope and fine structure cues, since periodicity cues may have been responsible for mediating speech perception, at least in part. Perhaps for this reason, several studies have lowpass filtered the Hilbert envelope (Swaminathan & Heinz, 2012) and thus removed or limited the contribution of the periodicity cues therein. The Hilbert transform has also been said to provide rather rapidly varying frequency information in the case of the TFS, compared to the original sub-band signals. Rectification and lowpass filtering is another technique that may be used in place of the Hilbert transform to extract E cues from the speech signal (Kong & Zeng, 2006). Rectification has the problem of not following the modulations in the E as faithfully as is possible with the Hilbert transform.

1.1 Need for the Study

The literature on the role of temporal cues in speech perception has spanned over 40 years and is ever expanding. Lorenzi & Moore (2007) has reviewed the role of E and TFS cues in speech perception in individuals with normal hearing and hearing impairment, and Shetty (2016) has done the same with respect to temporal cue enhancement for older adults. Given the large amount of data that has accumulated over the years since the publication of these articles, and the caveats in methodology and interpretation as has been outlined above, a synthesis of information is again needed to understand the relative contribution of E and TFS cues to speech perception. The present study thus aimed to look at the relative contribution of E and TFS cues in speech perception in normal hearing young adults. This study has not considered periodicity cues, since the number of articles comparing between all 3 cues was found to be fairly limited. The studies exploring the contribution of recovered E cues is also not included in the synthesis. This review specifically concentrates on how speech understanding is mediated by the "true" E and TFS. To facilitate unbiased comparison and reduce confounding variables, we have also limited our consideration to articles that have extracted E and TFS cues from the same speech material and with the same number of filter banks.

1.2 Aim of the Study

- To synthesize the evidence on speech perception abilities in adults with normal hearing when listening to speech stimuli having almost exclusively temporal information

1.3 Research Questions

- To document the importance of envelope cues in speech and lexical tone perception in adults with normal hearing, in quiet and adverse listening conditions

- To document the importance of fine structure cues in speech and lexical tone perception in adults with normal hearing, in quiet and adverse listening conditions

- To compare the relative importance of fine structure and envelope cues in speech and lexical tone perception in adults with normal hearing, in quiet and adverse listening conditions

Chapter 2

Methods

For this systematic review, the guidelines put forth in the preferred reporting items for systematic reviews and meta-analyses (PRISMA) (Page et al., 2021) were followed. A search strategy was developed and revised based on the relevance of the results generated, and these results were screened against a pre-set inclusion and exclusion criteria. Finally, data was extracted and risk of bias was assessed for each article in the final selection.

2.1 Search Strategy

The databases considered for the review were:

1. PubMed
2. ComDisDome
3. LLBA
4. IEEE Xplore
5. The American Institute of Physics (AIP)
6. Science Direct

Articles were searched for in each of the above, using a set of keywords. Due to restrictions on the number of keywords that may be used, as well as differences in the search and indexing algorithm in each of the databases, the keywords had to be modified and were, for the most part, unique for most databases (Table 2.1). Briefly, keywords related to two concepts (Temporal Cues and Speech Perception) were combined using Boolean operators; “OR” being used within keywords for each concept, and “AND” to combine across concepts.

Table 2.1

Keywords Used for Each Database

Database	Keywords
PubMed	("Temporal cue*"[tw] OR "temporal information"[tw] OR envelope[tw] OR ENV[tw] OR periodicity[tw] OR "fine structure"[tw] OR TFS[tw]) AND ("Speech Perception"[Mesh] OR "Speech Intelligibility"[Mesh] OR "speech perception"[tw] OR "speech intelligibility"[tw] OR "speech recognition"[tw] OR "speech segment*"[tw] OR "speech understanding"[tw] OR "speech identification"[tw] OR "speech discrimination"[tw] OR "speech reception"[tw] OR "consonant perception"[tw] OR "consonant discrimination"[tw] OR "vowel perception"[tw] OR "vowel discrimination"[tw] OR "sentence identification"[tw] OR "word identification"[tw] OR "phoneme identification"[tw] OR "syllable identification"[tw] OR "information transmission analysis"[tw])
ComDisDome	S1 AND (S2 OR S3) *S1:* (“Temporal cue*") OR "temporal information" OR envelope OR ENV OR periodicity OR (“fine structure”) OR TFS *S2:* "Speech perception" OR "speech intelligibility" OR "speech recognition" OR "speech segment*" OR "speech understanding" OR (“speech identification”) OR (“speech discrimination”) OR (“speech reception”) OR "consonant perception" OR (“consonant discrimination”) *S3:* "vowel perception" OR (“vowel discrimination”) OR (“sentence identification”) OR (“word identification”) OR (“phoneme identification”) OR (“syllable identification”) OR (“information transmission analysis”)

LLBA	S1 AND (S2 OR S3) *S1:* (“Temporal cue*") OR "temporal information" OR envelope OR ENV OR periodicity OR (“fine structure”) OR TFS *S2:* "Speech perception" OR "speech intelligibility" OR "speech recognition" OR "speech segment*" OR "speech understanding" OR (“speech identification”) OR (“speech discrimination”) OR (“speech reception”) OR "consonant perception" OR (“consonant discrimination”) *S3:* "vowel perception" OR (“vowel discrimination”) OR (“sentence identification”) OR (“word identification”) OR (“phoneme identification”) OR (“syllable identification”) OR (“information transmission analysis”)
IEEE Xplore	(“Temporal cue*” OR “temporal information” OR envelope OR ENV OR periodicity OR “fine structure” OR TFS) AND ("Speech perception" OR "speech intelligibility" OR "speech recognition" OR "speech segment*" OR "speech understanding" OR “speech identification” OR “speech discrimination” OR “speech reception” OR "consonant perception" OR “consonant discrimination” OR "vowel perception" OR “vowel discrimination” OR “sentence identification” OR “word identification” OR “phoneme identification” OR “syllable identification” OR “information transmission analysis”)
AIP	(“Temporal cue*" OR "temporal information" OR envelope OR ENV OR periodicity OR “fine structure” OR TFS) AND ("Speech perception" OR "speech intelligibility" OR "speech recognition" OR "speech segment*" OR "speech understanding" OR “speech identification” OR “speech discrimination” OR “speech reception” OR "consonant perception" OR “consonant discrimination” OR "vowel perception" OR “vowel discrimination”

	OR “sentence identification” OR “word identification” OR “phoneme identification” OR “syllable identification” OR “information transmission analysis”)
Science Direct	(“Temporal cue" OR "temporal information" OR envelope OR ENV OR periodicity OR “fine structure” OR TFS) AND ("Speech perception" OR "speech intelligibility")

No limits and filters were applied beforehand in any of the search engines. Snowballing/back referencing was carried out after the final list of articles was decided after full-text screening. This list was also screened in accordance with the PRISMA guidelines (Page et al., 2021).

2.2 Selection Criteria

Peer reviewed journal articles meeting the following inclusion criteria were selected for further analysis. The selection process followed the PRISMA flowchart (Page et al., 2021), as depicted below in the Results section in Figure 3.1. Two independent reviewers were involved in every stage of the selection process. Differences of opinion were resolved and the decision to remove or keep a particular article was done on the basis of mutual consensus. For the sake of uniformity, Hilbert’s (Hilbert, 1912) definition of E and TFS cues were adopted, and we have considered articles along the same lines. Articles discussing the periodicity cue were initially considered, but later discarded in favour of narrowing down the purview of this synthesis to E and TFS cues only.

2.2.1 Criteria for Inclusion

- Original articles from peer-reviewed journals
- Studies that considered E and TFS cues with reference to speech perception in quiet/noise
- Studies that have used behavioural methods in the assessment of speech perception (including, but not limited to, percent correct word recognition, speech reception thresholds and information transmission analysis)
- Studies that have used adult (age range: 18-45 years) humans with normal hearing (defined as hearing thresholds ≤20dBHL (BSA, 2004) in the ear under consideration) as participants

2.2.2 Criteria for Exclusion

- Studies that have used definitions of the E and TFS that are not in line with that of Hilbert
- Studies that have used only modelling techniques/neural/physiological estimates of speech perception
- Studies that have not extracted the E and TFS cues from the same speech material and the same number of frequency bands
- Studies based on animal participation
- Studies with ambiguous methodological procedures
- Review articles and case studies
- Publications in languages other than English

2.3 Data Extraction

Data was extracted into a non-standardized format that took into account the possible important variables of the studies. This included the research question of the study, methodological variables such as participant demographics and details of the speech stimuli, processing and testing conditions, and the results and conclusions of each of the studies.

A modified version of the methodological quality appraisal tool as was used by Gunjawate et al. (2018) (who had in turn modified from Downs and Black (1998) and Sanderson et al. (2007)) was used to assess the quality of all studies considered for the review. The questions in the modified appraisal tool are provided below:

Q1. Was the aim/objective of the study clearly defined?

Q2. Were the participant inclusion criteria clearly described?

Q3. Are the main study findings explained?

Q4. Are the main outcome measures clearly stated?

Q5. Were the investigators blinded to the participant characteristics to reduce bias?

Q6. Is there a clarification for the appropriateness of the sample size studied?

Q7. Have the investigators provided a clarification about the settings under which the findings can be applied?

Chapter 3

Results

The PRISMA chart given below in Figure 3.1 depicts the flow of information in this systematic review. The search strategies yielded a total of 9,138 studies across all 6 databases, as shown in Table 3.1. These were imported and pooled in the Zotero Library. Duplicate screening carried out by the reference manager resulted in 7,997 articles. After title and abstract screening, 40 articles remained to be assessed in full-text. Of these, 20 articles were removed, as they did not meet the inclusion criteria. The eliminated articles used a different definition for the TFS cue, compared the TFS with only the recovered E cues, did not extract the E and TFS from the same number of frequency bands, and/or did not specify the age of the participants. Hand-searching the references of the final selected articles revealed 4 more articles that fit the inclusion criteria, and thus a total of 24 articles were included in the systematic review. A summary of the twenty-four finalized articles is provided in Table 3.2.

Table 3.1

Results from the Databases Included in the Review

Database	**Date of Last Search**	**No. of Results**
PubMed	11/01/21	1,371
ComDisDome	08/01/21	729
LLBA	14/01/21	1,614
IEEE Xplore	14/01/21	1,929
AIP	14/01/21	79
Science Direct	09/01/21	3,416

Figure 3.1

PRISMA Flow Chart Showing Flow of Information (Page et al., 2021)

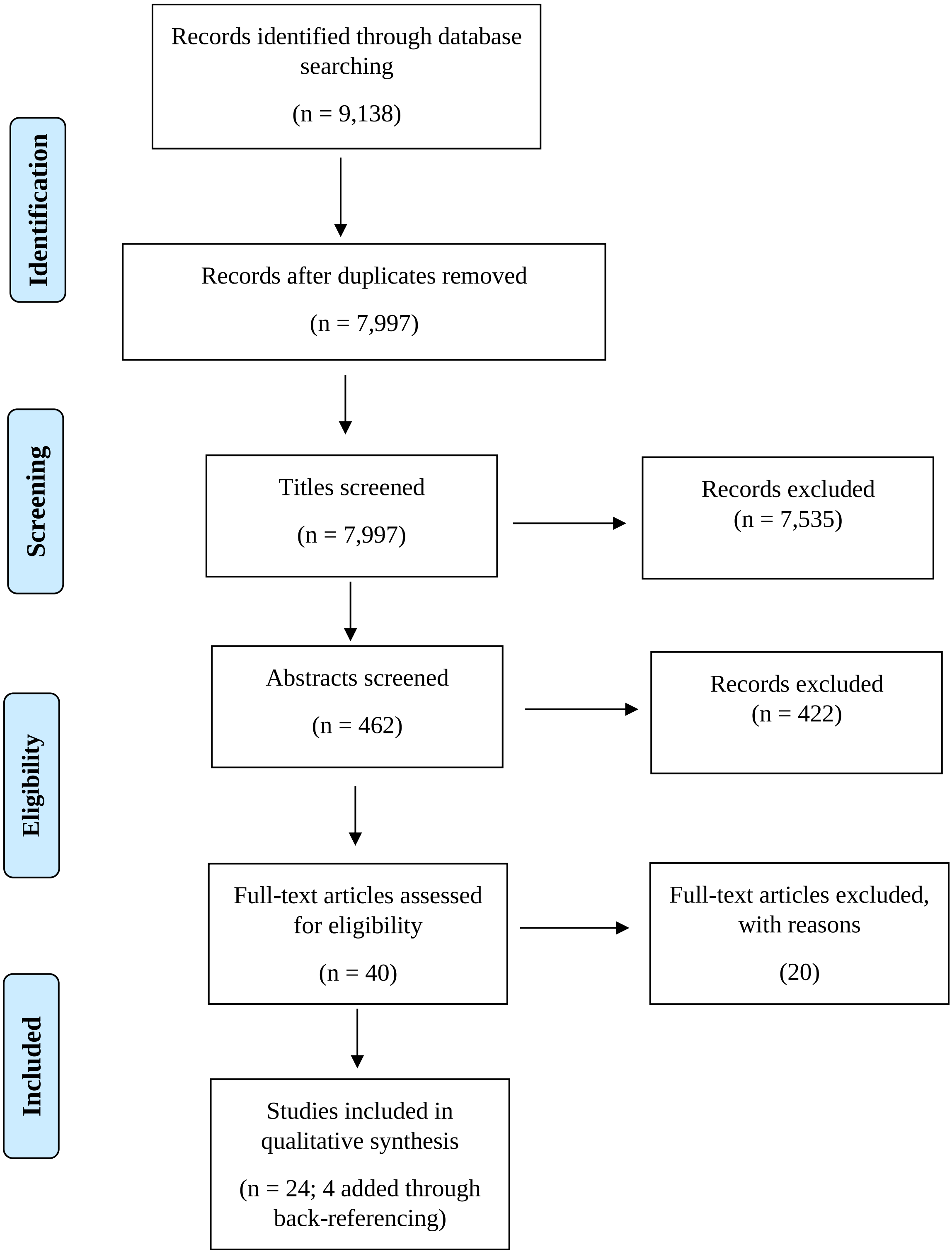
Identification
Screening
Eligibility
Included
Records identified through database searching
(n = 9,138)
Records after duplicates removed
(n = 7,997)
Titles screened
(n = 7,997)
Records excluded
(n = 7,535)
Abstracts screened
(n = 462)
Records excluded
(n = 422)
Full-text articles assessed for eligibility
(n = 40)
Full-text articles excluded, with reasons
(20)
Studies included in qualitative synthesis
(n = 24; 4 added through back-referencing)

Table 3.2

Summary of the Articles Included in the Review

1. Speech Perception in Quiet Listening Conditions

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
1.	***Speech perception problems of the hearing impaired reflect inability to use temporal fine structure*** (Lorenzi et al., 2006)	To investigate the ability of individuals with sensorineural hearing loss to understand speech in background noise (and hence compared the ability of individuals with normal hearing and those with sensorineural hearing loss to understand intact speech and speech processed to remove E or TFS cues)	Experimental Study	**Subjects** • Number: 7 • Age Range: Mean = 26 years; Range = 21-35 years • Gender: Not specified • Language: Not specified • Ear: Monoaural, right ear • Control condition: None **Stimuli** • Speech Material: 3 tokens each of 16 /aCa/ disyllables amounting to a total of 48 tokens, pronounced by a female French speaker • Conditions: Intact, E and TFS conditions • Noise: None	**Results** • Intact Speech: 100% scores • E Speech: >90%, after training • TFS Speech: close to 90%, after more training than for E speech **Conclusions** Although E speech important for speech perception in quiet, TFS speech by itself also can provide high intelligibility in similar situations

- Processing used: Hilbert transform: Intact: summed over all bands; E stimuli were processed using zero-phase, 6th order Butterworth filters and low pass filtered at 64Hz and tone-vocoded; TFS stimuli were power-weighted.
- No. of analysis bands: 16 adjacent 0.35-octave bands
- Filter Bank Range: 80-8020Hz
- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: Under monoaural (right ear) conditions with Sennheiser HD25-1 headphones; setting not specified
- Level: 75dBA
- Familiarization: For each condition, with 5-minute sessions until the responses stabilized at less than 9% change in scores across 4 consecutive sessions. E and

TFS conditions were interleaved and the condition to attain stability first tested first, the other condition trained until stability achieved; TFS: 10-17 sessions; E: 4-12; Intact: 4-6

- Testing: Intact condition was tested first, then the E and TFS speech together in an interleaved fashion; the 48 tokens were presented in random order
- Response mode: Not specified; no feedback provided
- Outcome measure & Scoring: Percent correct identification

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
2.	***Speech identification based on temporal fine structure cues.*** (Sheft et al., 2008)	To determine if the speech information obtained from temporal fine structure cues (TFS) is due to the recovered envelope or due to TFS in and of itself	Experimental Study	**Experiment 1:** **Subjects** • Number: 7 • Age Range: 21-32 years (mean: 24, SD: 4 years) • Gender: Not specified • Language: Native French speakers • Ear: Monoaural, Right ear • Control condition: None **Stimuli** • Speech Material: 48 syllables, 3 exemplars each of 16 /aCa/ materials pronounced by a female French speaker • Conditions: 3 TFS and 1 E conditions • Noise: None • Processing used: Hilbert transform; TFS: 2 phase (PMz, PMr) modulated conditions with the second	**Results** Experiment 1: Word Identification Scores • *E condition:* Minimal effect of training; even with just 4 sessions, mean identification score of 90% achieved • *TFS condition:* Individual scores varied between 50-90% correct across conditions for the two best sessions. Performance improved with training across the 10 sessions; PMz: 80%, PMr: 70%, FM: 65% Information Transmission Analysis • *E condition:* Greatest information transmitted was for voicing and nasality, as was for the TFS conditions, however lesser information was received for manner and least for place of articulation, unlike the TFS conditions.

	one having random carrier starting phase, and 1 frequency modulated (FM) condition with the frequency deviations limited to within the analysis filter bandwidth, all TFS conditions rms power-weighted; E: envelopes extracted lowpass filtered at 64Hz using a third order zero phase Butterworth filter, tone-vocoded • No. of analysis bands: 16 contiguous 0.4-octave bands • Filter Bank Range: 80-8020Hz • Transition of overlap: Not specified **Procedure** • Test setting & Transducer: Double walled sound proof booth, under Sennheiser HD212 Pro, right ear • Level: 80dBA • Familiarization: Within the testing itself • Testing: TFS:10 sessions for each of the 4 conditions;	• *TFS condition:* Information transmitted improved with training for all 3 conditions. Information transmitted was more overall for nasality and voicing, then place, and least for manner of articulation. Across conditions, more information was transmitted for the PMz then PMr then FM conditions. Experiment 2: Word Identification Scores • Overall, identification scores did not show any significant reduction compared to the higher presentation level of 80dB(A), for any of the 4 processing conditions. Information Transmission Analysis • Significant reduction with level for all processing conditions; maximum reduction was observed for manner cues in E

conditions interleaved and the stimulus order randomized; E: 4 sessions following testing with TFS stimuli; 16 consonantal choices were displayed on a computer monitor

- Response mode: Mouse click on the correct option
- Outcome measure & Scoring: Percent correct identification; information transmission analysis

Experiment 2:
Subjects

- Number: 7
- Age Range: 21-32 years (mean: 24, SD: 4 years)
- Gender: Not specified
- Language: Native French speakers
- Ear: Monoaural, Right ear
- Control condition: None

condition and place cues in the PMz condition.

Conclusions

- Greater than moderate levels of consonant perception possible through the use of TFS cues
- E and TFS cues convey phonetically different information; E cues convey more information on manner than on place, and vice-versa for TFS cues

Stimuli

- Speech Material: 48 syllables, 3 exemplars each of 16 /aCa/ materials pronounced by a female French speaker
- Conditions: 3 TFS and 1 E conditions
- Noise: None
- Processing used: Hilbert transform; TFS: 2 phase (PMz, PMr) modulated conditions with the second one having random carrier starting phase, and 1 frequency modulated (FM) condition with the frequency deviations limited to within the analysis filter bandwidth, all TFS conditions rms power-weighted; E: envelopes extracted lowpass filtered at 64Hz using a third order zero phase Butterworth filter, tone-vocoded
- No. of analysis bands: 16 contiguous 0.4-octave bands
- Filter Bank Range: 80-8020Hz

- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: Double walled sound proof booth, under Sennheiser HD212 Pro, right ear
- Level: Average level of 45dBSPL
- Familiarization: Within the testing itself
- Testing: The 3 TFS and 1 E conditions were tested over 4 sessions each, only one session for the unprocessed low-level speech; order of presentation of the 4 processed conditions was interleaved and randomized; 16 consonantal choices were displayed on a computer monitor
- Response mode: Mouse click on the correct option
- Outcome measure & Scoring: Percent correct identification;

information transmission analysis

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
3.	***Discrimination of speech sounds based upon temporal envelope versus fine structure cues in 5- to 7-year-old children.*** (Bertoncini et al., 2009) **Data pooled for adults and children not considered, hence information transmission analysis data omitted***	To determine the ability of children and adults to discriminate speech on the basis of E and/or TFS cues	Experimental Study	**Subjects** • Number: 10 • Age Range: M=23 years; SD=2 years • Gender: Not specified • Language: French speakers • Ear: Binaural • Control condition: None **Stimuli** • Speech Material: 3 tokens each of 5 /aCa/ disyllables spoken by a female French speaker using clear speech • Conditions: Intact, E and TFS conditions • Noise: None • Processing used: (Acc. to Gilbert & Lorenzi (2006)) Hilbert transform to produce the E and TFS stimuli; Intact: summed over all bands; E stimuli were processed using zero-phase, 6th order	**Results** Response Accuracy Intact & E conditions: d' value of around 2.8-3.0 TFS: d' value around 1.5 E/TFS order: d' difference E-Intact: around 0.3 TFS-Intact: around -1 TFS/E order: d' difference E-Intact: around -0.1 TFS-Intact: around -1.5 Response Latency E/TFS order: Latency difference E-Intact: around 100ms TFS-Intact: around 400ms TFS/E order: Latency difference E-Intact: around 20ms TFS-Intact: around 400ms **Conclusions** Discrimination abilities better and latencies shorter with E than TFS stimuli

Butterworth filters and low pass filtered at 64Hz and tone-vocoded; TFS stimuli were power-weighted

- No. of analysis bands: 16 contiguous 0.35-octave bands
- Filter Bank Range: 80-8020Hz
- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: Sound treated room, under Sennheiser HD-580 (binaural)
- Level: 75dBA
- Familiarization: None
- Testing: VRSID paradigm; the subject sat facing a computer monitor which had a background image; background stimulus of /aba/ and 20 (5 trials x 20 consonants) change/5 no-change trials (ISI: 450-1200); half the subjects were tested in the order E/TFS, the other half in TFS/E

- Response mode: Press button if there is a change in stimuli. Visual feedback provided for both correct and incorrect (misses/false alarms) responses
- Outcome measure & Scoring: Discrimination responses (d’) for response accuracy; information transmission analysis; response latency measures

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
4.	***The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences.*** (Fogerty & Humes, 2012)	To investigate the cues responsible for the perception of vowels and consonants in the context of words and sentences	Experimental Study	**Subjects** • Number: 14 (in groups of 2) • Age Range: 19-23 years (M=21 years) • Gender: Not specified • Language: Native American English • Ear: Monoaural; right ear • Control condition: None **Stimuli** • Speech Material: Sentence and word material; 42 sentence materials taken from the TIMIT database, each half of the set spoken by a different male and female talker in the North Midland dialect; 148 /CVC/ words taken from recordings by Takayanagi et al. (2002) in a male voice of General American dialect and at two levels of difficulty	**Results** • Sentences ○ V only sentences better than C only sentences ○ TFS group performed significantly better in the V only condition; E group in the C only condition • Words ○ TFS group better for V only and for E group for C only words • Across stimuli, performance better for V only sentences than words

	• Conditions: Each group of subjects were assigned to an E only or a TFS only condition, in each group: 2 speaker (male vs female) x 2 stimuli (word vs sentence) x 2 segmental condition (vowel (V) only vs consonant (C) only) • Noise: For noise replacement, the noise was matched to the LTAS of the concatenation of all sentences/words and at -16dBSNR relative to the sentences; noise for the E only/TFS only stimuli also matched the LTAS of each original sentence sample • Processing used: Hilbert transform; for E only speech, E was at 11dBSNR and TFS at -5dBSNR; and vice versa for TFS speech • No. of analysis bands: 3 bands, at equal distances on the basilar membrane • Filter Bank Range: 80-6400Hz	**Conclusions** • TFS cues present only in the vowel conditions in words and sentences • E cues present in both vowel and consonant condition in both contexts

- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: In a sound-proof booth, under ER-3A insert earphones
- Level: 70dBSPL
- Familiarization: Provided with other word and sentence stimuli than the ones used for testing
- Testing: The E group and TFS group were tested on word and sentence materials in V only and C only conditions
- Response mode: Sentences: repeat the sentence heard; /CVC/: type out word on a computer
- Outcome measure & Scoring: Percent correct scores; RAU; all words in sentence were scored, should be exact to be scored as correct

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
5.	***Predictions of Speech Chimaera Intelligibility Using Auditory Nerve Mean-Rate and Spike-Timing Neural Cues.*** (Wirtzfeld et al., 2017)	To investigate the neural mean rate and spike timing cues to phoneme perception	Experimental Study	**Subjects** • Number: 5 • Age Range: 18-21 years • Gender: Not specified • Language: Native speakers of English • Ear: Not specified • Control condition: None **Stimuli** • Speech Material: 50 NU-6 words with carrier phrase, spoken by a male American English speaker; total 1750 tokens after processing • Conditions: 5 conditions; Speech ENV with WGN TFS; Speech ENV with WN TFS, Speech TFS with WGN ENV, Speech TFS with MN ENV, Speech TFS with flat ENV • Noise: White Gaussian Noise (WGN) and Matched Noise (MN)	**Results** • With ENV speech, performance increased with number of bands, the reverse was true for TFS speech • For number of filters <6, ENV cues helped consonant recognition more than that of vowels, the reverse for TFS in most cases • >6 filters, performance saturates for both consonants and vowels for ENV speech • Phoneme recognition better for ENV with MN TFS than WGN TFS for filters <6 • Speech TFS with MN ENV performance worse than WGN ENV and Flat ENV **Conclusions** Envelope major cue for speech perception than fine structure, especially with increasing number of bands

- Processing used: Hilbert transform; E not LP filtered; chimerizer
- No. of analysis bands: 1, 2, 3, 6, 8, 16, 32
- Filter Bank Range: 80-8820Hz
- Transition of overlap: 25% pf the BW of the narrowest filter

Procedure

- Test setting & Transducer: Quiet room, under Sennheiser HDA 200 headphones
- Level: 65dBSPL
- Familiarization: None
- Testing: 1 hour for each condition, order of tokens randomized within each listener; each condition, 7 frequency bands, 50 test words each
- Response mode: Repeat word
- Outcome measure & Scoring: Percent correct phoneme recognition; phonemic

scoring main method of scoring

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
6.	***Role of short-time acoustic temporal fine structure cues in sentence recognition for normal-hearing listeners.*** (Hou & Xiu, 2018)	To assess the relative contribution of E and TFS cues to sentence recognition in short segments; evaluate the benefit of using short segments in studying speech perception	Experimental Study	**Subjects** • Number: 52 • Age Range: 18-31 years (M= 24.6 years) • Gender: 21 male, 31 female • Language: Native speakers of English • Ear: Not specified • Control condition: None **Stimuli** • Speech Material: 33 lists of AzBio English corpus, 20 sentences each; 140 words per list, spoken by 2 male and 2 female talkers • Conditions: 4 experiments; Flattened TFS, Short term TFS, Randomized TFS, Constant frequency TFS; 18 conditions (3 channels x 18 segments) • Noise: None	**Results** • FTFS: ○ Mean scores across segment durations: ○ 9, 45 and 88% for 1, 2 and 4 ERBs, respectively • STFS: ○ Scores >90% at 4ERB BW ○ 1, 2 ERBs at 300ms, ~100% at 50ms • RTFS, CTFS: ○ Similar results; 50ms: E at 20Hz, score of ~100% with 32 bands but ~50% with 8 bands ○ Further decrease for 100 and 150ms

	- Processing used: Hilbert transform; sentences segmented into 50, 100, 150, 200, 250 or 300ms; FTFS: E replaced by 1, only TFS; RTFS: E preserved at rms value of each segment, TFS unaltered; RTFS: TFS removed by randomizing phase; CTFS: instantaneous phase replaced with sinusoids - No. of analysis bands: 8, 16 and 32 (4, 2 and 1-ERB wide, respectively) - Filter Bank Range: 64-8932Hz - Transition of overlap: Not specified **Procedure** - Test setting & Transducer: Sound booth, under supra-aural earphones - Level: Not specified - Familiarization: With 72 processed sentences; feedback provided	**Conclusions** Short term TFS cues, along with E cues, help speech perception for normal hearing individuals in quiet

- Testing: Order of the 18 conditions randomized; 360 sentences in total
- Response mode: Transcription of entire sentence
- Outcome measure & Scoring: Percent correct sentence recognition; all words included for scoring

2. Speech Perception in Adverse Listening Conditions

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Methods	Findings of the Study
7.	***Temporal envelope and fine structure cues for speech intelligibility*** (Drullman, 1995)	Main: Relation between MTF, speech intelligibility, compression and noise Also compared intelligibility with preserved/altered envelope and fine structure cues	Experimental Study	**Subjects** - Number: 60 subjects, divided into 5 groups, each group allotted to a certain processing condition - Age Range: 18-30 years - Gender: Not specified - Language: Not specified - Ear: Monoaural - Control condition: None **Stimuli** - Speech Material: 130 Dutch sentences of 8-9 syllables, spoken by a female speaker - Conditions: 6 processing conditions:	**Results** - REF – poorer scores compared to SN and FT for all target conditions - SRT for REF, SN and FT: 5.5, 6.5 and 12dB resp. - Intact fine structure in the presence of a noisy envelope provides 1 dB betterment in SRT - Noisy envelope more detrimental to intelligibility than envelope with flattened troughs, with same speech fine structure - NFS – mean score 98.3% - high intelligibility – hence envelope major cue

- SN –
  - Fine structure: speech
  - Envelope – Speech + noise
- FT –
  - Fine structure: speech
  - Envelope: Speech in peaks, flat troughs
- FP –
  - Fine structure: speech
  - Envelope: Speech in troughs, flat peaks
- BLK –
  - levels above target level made equal to Leq, levels below fixed to 0
- REF –
  - Fine structure: speech + noise
  - Envelope: speech + noise
- NFS –
  - Fine structure: noise
  - Envelope: speech

SN, FT, FP, BLK – 12 target conditions; REF – 6

<u>Conclusions</u>-

- Envelope cue more important than fine structure, which provides minimal cues for speech intelligibility
- In processing with 24 ¼ octave bands, intact envelope and noisy fine structure – high intelligibility; intact fine structure and noisy envelope – only average score of 17%

- Noise: Matched the LTAS of the 130 sentences
- Processing used: Analysis-resynthesis algorithm; Hilbert transform;

  - "Target level" of each ¼ octave band: an imaginary line passing through the envelope and expressed in dB with respect to a reference of 0dB, that being the long term rms level (Leq) based on the 130 sentences

- No. of analysis bands: 24 ¼ octave bands; linear-phase FIR filters, min. 80dB/octave
- Filter Bank Range: 100-6400Hz
- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: Sound-proof room, under Sony MDR-CD999, at ear of preference
- Level: Approx. 65dBA

- Familiarization: With a list of 11 sentences in a representative condition before testing
- Testing:
  - The 130 sentences were divided into 5 lists of 12; first stimuli of list 1 and 2 were also used as first stimuli for lists 11 and 12
  - Sequences for each processing and target condition was made according to 12x12 Latin square (SN, FP, FT, BLK), and 6x6 for REF
  - Each subject of the respective groups for SN, FP, FT, BLK was presented with each sequence; one sequence for 2 subjects in REF group
  - NFS: 4 separate lists of 11 sentences (male voice) presented to 3 subjects
- Response mode: Instructed to repeat the sentences as accurately as possible
- Outcome measure & Scoring: Percent correct score; scored

as correct only if entire sentence repeated correctly; RAU

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
8.	***Effects of periodic interruptions on the intelligibility of speech based on temporal fine-structure or envelope cues.*** (Gilbert et al., 2007)	To find out if the temporal envelope (E) and fine structure (TFS) both carry identical speech information	Experimental Study	**Subjects** • Number: 7 • Age Range: 23 years (S.D.: 5 years) • Gender: Not specified • Language: Native French speakers • Ear: Diotic presentation • Control condition: None **Stimuli** • Speech Material: Three tokens each of 16 /aCa/ syllables spoken by a female talker, total of 48 stimuli • Conditions: Intact (containing both E and TFS information), E only and TFS only; periodic interruptions-sinusoidal and square waves • Noise: None • Processing used: Hilbert transform to derive the E and	**Results** • The scores across all conditions were greater than chance (6.25%) and >66% • Across conditions, E stimuli was better than TFS by 13.5 percentage points The square wave interruptions were more disruptive to speech intelligibility than sinusoidal interruptions, but at low modulation frequencies only • The disruptive effect was more for E than TFS information (from 2 to 64Hz, the drop was by 15 and 24 percentage points for TFS and E information, respectively) **Conclusions** • Both E and TFS are robust in presence of periodic interruptions

	TFS information; Intact: summed over all bands; E-LPF zero-phase, sixth order Butterworth, tone-vocoded; periodic interruptions-sinusoidal and square waves of 50% duty cycle, 100% depth and at rates from 2 through 64Hz in octave steps • No. of analysis bands: 16 contiguous 0.4 octave bands • Filter Bank Range: 80-8020Hz • Transition of overlap: Not specified **Procedure** • Test setting & Transducer: Sound proof booth under Sennheiser HD565 headphones • Level: 70dBA • Familiarization: Trained until greater than 90% correct performance was achieved for E (4-11 sessions) and TFS (9-29 sessions) stimuli (final mean scores were 4%	• In presence of periodic interruptions, E stimuli were more affected than TFS, and this disruption occurred at lower modulation frequencies; greater modulation masking effect for E vs TFS stimuli • Hence E and TFS stimuli do not convey identical speech information

higher for E than TFS stimuli)

- Testing: 4 identical set of the 48 /aCa/ stimuli were presented at random for each condition. The 16 consonantal choices were displayed on a computer monitor
- Response mode: Mouse click on the correct response. No feedback was provided
- Outcome measure & Scoring: Percent correct identification

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
9.	***Abnormal processing of temporal fine structure in speech for frequencies where absolute thresholds are normal.*** (Lorenzi et al., 2009)	To investigate the speech perception abilities of hearing-impaired listeners with stimuli spectrally limited to regions of normal thresholds (Additionally, tested the same with individuals with normal hearing)	Experimental Study	**Subjects** • Number: 12 • Age Range: 21-46 years (Mean= 29 years) • Gender: Not specified • Language: Not specified • Ear: (?)Binaural • Control condition: None **Stimuli** • Speech Material: 3 tokens each of 16 /aCa/ disyllables spoken by a female French speaker (48 stimuli in total) • Conditions: Intact, E and TFS • Noise: None • Processing used: Hilbert transform: Intact: summed over all bands; E stimuli were processed using zero-phase, 6th order Butterworth filters and low pass filtered at 64Hz and tone-vocoded; TFS	**Results** *Missing values for the 3 subjects who did not return for the E testing* Intact speech: >60% E speech: around 60-80% TFS speech: 20-50% RAU transformed scores as E/Intact: 88% TFS/Intact: 44% With transition band masked, significant effect only on E speech (12 percentage point decrease) **Conclusions** • Significant amount of TFS information is available above 1.5kHz, compared to E information

	stimuli were power-weighted. Resultant stimuli lowpass filtered at 1.5kHz (-72dB slope, Butterworth filter) • No. of analysis bands: 16 contiguous 0.35-octave bands • Filter Bank Range: 80-8020Hz • Transition of overlap: Not specified **Procedure** • Test setting & Transducer: Under Sennheiser HD580 earphones; setting not specified • Level: 75dBA • Familiarization: For 6 5-minute sessions • Testing: For 4 sessions; Intact and TFS stimuli presented first, subjects called later for E speech testing (9/12) • Response mode: Not specified	• Relative importance of E over TFS information in speech perception

- Outcome measure & Scoring:
  Percent correct identification;
  RAU transformed scores

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
10.	***Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine structure or envelope cues*** (Ardoint & Lorenzi, 2010)	To investigate if E and TFS cue distinct phonetic information	Experimental Study	**Subjects** • Number: 7 • Age Range: 20-24 years • Gender: Not specified • Language: Native French speakers • Ear: Monoaural • Control condition: None **Stimuli** • Speech Material: Three tokens each of 16 /VCV/ nonsense syllables spoken by a male (F0: 115Hz) and female (F0: 221Hz) speaker; 48 tokens each • Conditions: 2 speakers x Intact, E and TFS conditions x highpass and lowpass conditions x 6 cutoff frequency conditions • Noise: None • Processing used: Hilbert transform; Intact: summed	**Results** Mean Identification Scores Intact: 99% E: 94-99% TFS: 70-75% • Lowpass filtering condition: for both E and TFS and all cutoff frequencies, significantly above chance • Highpass filtering condition: Significantly above chance for ○ TFS: Only below 2542Hz ○ E: upto 6030Hz Crossover frequencies • Not significantly different between E and TFS speech or

	over all bands; E: third order zero phase Butterworth filter, lowpass cutoff frequency of 64Hz, modulated sinusoidal carriers and tone-vocoded; TFS generated as in the phase modulated condition (Sheft at al., 2008), power weighted; cutoff frequencies for high and lowpass filtering conditions: 254, 803, 1429, 2542, 3390, or 6030 Hz • No. of analysis bands: 16 adjacent 0.35-octave bands • Filter Bank Range: 80-8020Hz • Transition of overlap: Not specified **Procedure** • Test setting & Transducer: Setting as such not specified; under Sennheiser HD 212 Pro earphones monoaurally • Level: 80dBSPL • Familiarization: Training sessions were provided using	across speakers (ranged between 1.3-1.6Hz) Gradients • No significant difference across speakers • Male speaker: significant difference between E and TFS conditions for lowpass filtering conditions between 803-2542Hz • Highpass filtered Condition; significantly different from 0 over ○ E: 803–6030 Hz; ○ TFS: 1429–2542Hz • Lowpass filtered Condition; significantly different from 0 over ○ E and TFS: 254– 2542 Hz

	unfiltered E (at least 3 sessions) and TFS (at least 5 sessions); training ended when stable scores achieved over 3 consecutive sessions • <u>Testing:</u> Around 20h for each listener; the speaker, the processing and filtering conditions and cutoff frequencies were randomized across the listeners (speaker and processing conditions were tested separately as whole blocks) • <u>Response mode</u>: Identify the presented /VCV/ from the 16 options on a computer monitor; no feedback was provided • <u>Outcome measure & Scoring:</u> Percent correct identification across crossover frequencies and gradients	✓ No speaker difference for crossover frequencies and gradients ✓ For E and TFS speech, similar crossover frequencies, ranged between 1.3-1.6Hz ✓ Gradients significantly different for E and TFS speech for various filtering conditions **<u>Conclusions</u>** • E cues more effective than TFS, however both required for speech perception • E and TFS cues convey complementary phonetic distinctions between 1 and 2.5kHz • E cues convey speech information upto 6kHz, unlike TFS cues

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
11.	***Role of spectral and temporal cues in restoring missing speech information.*** (Gilbert & Lorenzi, 2010)	To investigate the contribution of spectral, E and TFS cues to perception of interrupted speech	Experimental Study	**Subjects** • Number: 32; divided into 4 groups • Age Range: M=21.9 years; SD=2 years • Gender: Not specified • Language: Not specified • Ear: Monoaural; right ear • Control condition: None **Stimuli** • Speech Material: Meaningful sentence materials: 16 lists of 8 sentences each, each spoken by a male French speaker • Conditions: 4 stimulus conditions; Reference "REF", Partly Empty "PEMP", Vocoded "VOC" and Partly Vocoded "PVOC", used in control condition. In the main identification experiment,	**Results** Control Conditions 100% across the 4 conditions and across the groups Main Identification Condition (mean scores) REF: 58% PEMP: 46% VOC: 31% PVOC: 38% Three of the four groups showed the trend REF>PEMP>PVOC>VOC, the fourth group showed comparable scores across REF, PEMP and PVOC With only TFS cues (REF), performance was reduced by 42 percentage points; with only E cues (VOC), this decrease was by 69 percentage points Unexpectedly, PVOC lower in scores than PEMP- limited contribution of the 21 vocoded bands to speech perception

each of these conditions were interrupted as well, making a total of 8 different conditions

- Noise: None
- Processing used: REF: No further processing after being passed through the gammatone filters; PEMP: 21 filters replaced by 0s; VOC: Full-wave rectification, lowpass filtering through a 7500th FIR filter, cutoff f: ERBn/2, modulated spectrally flat noises; PVOC: 11 filtered, 21 filtered then vocoded outputs; interrupted conditions used 120ms silent gap, 50% duty cycle and square waves
- No. of analysis bands: 32 gammatone filters
- Filter Bank Range: 100-9106Hz
- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: Double-walled, sound proof

Conclusions

- Degradation of TFS cues reduces the ability to identify interrupted speech
- The role played by E cues are not clear, however, TFS cues may be important in such identification in the presence of E cues, and not by itself

booth, under Sennheiser HD25 1 II (right ear)

- Level: 62dBSPL (continuous PEMP); 65dBSPL (continuous, VOC, PVOC, REF)
- Familiarization: With a list of 24 sentences, each under all 8 conditions
- Testing: Each group of subjects were tested with a control and a main experiment; in the control experiment, 4 lists presented, each corresponding to one control condition; in the main experiment, remaining 12 lists used, with these lists divided into 4 groups of 3 lists each and each group corresponding to one of the interrupted processing conditions; Latin square design used
- Response mode: Mode as such not specified
- Outcome measure & Scoring: Percent correct identification; keywords identified for each sentence; 23-26 in control

and 70-76 in the main identification experiment

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
12.	***Benefit of temporal fine structure to speech perception in noise measured with controlled temporal envelopes*** (Eaves et al., 2011)	To investigate whether the benefit to speech perception in vocoded speech with TFS cues is due to the modifications inadvertently created in the temporal envelope during processing	Experimental Study	**Subjects** • Number: 40 (16 in main, rest in control experiment) • Age Range: 18-42 years (M=21.3 years) • Gender: Not specified • Language: Native English speakers • Ear: Binaural • Control condition: First: 16 subjects, IHR sentences, quiet, and noise, 6 conditions: 2 processing x 3 noise (steady, modulated at 30dB, quiet), 60dbA; Second: 4 subjects, IEEE, 60 & 65 dBA, 8 conditions: 2 presentation levels, 2 processing, 2 noise (steady, 30dB modulated), Third: 4 subjects, IEEE, 6 conditions: 2 processing, 3 noise (steady, 30dB and 60dB modulated), 60dBA	**Results** Main experiment • Lower SRTs: ○ IHR than IEEE sentences by 0.6dB ○ Modulated than steady noise by 8.2dB ○ ENV&TFS than ENV by 2.4dB ○ Once-BPF than twice-BPF stimuli by 0.2dB • TFS contributed in modulated noise significantly better than steady (3.0 vs 1.9dB) Control experiments • Revealed that audibility of sentences and floor effects did not influence results

Stimuli	**Conclusions**
• Speech Material: Two sets of sentence stimuli (IHR & IEEE) • Conditions: 2X2X2X2 factorial condition (2 conditions each of stimuli, noise (modulated, steady), processing (ENV, ENV&TFS), filtering (once-BPF and twice-BPF) • Noise: Produced by combining sinewaves and shaped to LTAS of each sentence • Processing used: Hilbert transform; ENV extracted and tone-vocoded; band pass filtered once (once-BPF) or twice (twice-BPF), according to the condition; 37 SNRs from -36 to +36 in 2dB steps; for negative SNRs, stimulus level reduced, noise reduced in positive • No. of analysis bands: 32 FIR filters • Filter Bank Range: 100-10,000 (the 6dB down points	Inclusion of TFS cues enhances the intelligibility of sentences in noise, an effect not caused due to modifications of E cues during processing.

on the low and high frequency sides, respectively)

- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: Setting as such not specified; under Sennheiser HD580 headphones, binaurally
- Level: 60dBA
- Familiarization: 15 trials in each condition
- Testing: Lists of 30 sentences, counterbalancing using Williams-square, lists presented equal number of times per condition, SNR controlled adaptively
- Response mode: Repeat the sentence as accurately as possible
- Outcome measure & Scoring: SRT50; IHR: 3 keywords repeated, IEEE: >3 for full score; Probit units

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
13.	***Perceptual weighting of individual and concurrent cues for sentence intelligibility: Frequency, envelope, and fine structure*** (Fogerty. 2011)	To measure the perceptual weights given to E and TFS cues and to different spectral regions, for speech perception	Experimental Study	**Experiment 1** **Subjects** • Number: 8 • Age Range: 19-23 years • Gender: Not specified • Language: Native speakers of American English • Ear: Monoaural (right) • Control condition: None **Stimuli** • Speech Material: 240 sentences from the IEEE database, spoken by a male speaker • Conditions: 10 conditions; 2 processing conditions (E and TFS) x 5 SNRs (11, 5, 2, -1, -7) • Noise: Speech shaped noise matched to the power spectrum of each sentence	**Results** Experiment 1 • Difference between relative weights for E and TFS cues across all 3 bands insignificant Experiment 2 • E2 and E3 channels were more perceptually weighted, and less on TFS1 and TFS3 • With respect to the second half of the sentence, less weight given to E1 and more to TFS3 **Conclusions** • When more than one cue is available for phonetic distinction, listeners place equal weight on E and TFS cues in the midfrequency range (528-1941Hz)

	• Processing used: Hilbert transform; E and TFS were extracted from sentences that were combined with noise to create an SNR of -5dB; for the test sentences, non-target portion was at -5dBSNR while the target portions were scaled across a range of SNRs from 11 to -7 • No. of analysis bands: 3 bands • Filter Bank Range: Band 1: 80–528Hz; Band 2: 528–1941Hz; Band 3: 1941–6400Hz • Transition of overlap: Not specified **Procedure** • Test setting & Transducer: Sound attenuating booth; under ER-3A insert earphones, unilaterally (right ear) • Level: Maintained at 70dBSPL with calibration	• Across spectral regions, E cues given more importance • Little perceptual weight provided to low frequency regions for both E and TFS in continuous speech

- Familiarization: Using male voice stimuli from the TIMIT database, in the corresponding conditions as used in the experiment; no feedback provided
- Testing: 24 sentences, randomized (120 keywords) presented for each of the 10 conditions
- Response mode: Repeating the sentence as accurately as possible; no feedback provided
- Outcome measure & Scoring: Percent correct identification; keywords scored as correct or incorrect if produced accurately, regardless of order; point biserial correlation measured for all listeners

Experiment 2
Subjects

- Number: 10 (separate)
- Age Range: 18-26 years
- Gender: Not specified

- Language: Native speakers of American English
- Ear: Monoaural (right)
- Control condition: N/A

Stimuli

- Speech Material: 600 sentences from the IEEE database
- Conditions: 6 acoustic information “channels”, 3 bands x 2 processing conditions (E and TFS)
- Noise: Speech shaped noise matched to the power spectrum of each sentence
- Processing used: Hilbert transform; the SNR in each of these 6 channels were varied independently in 3dB steps from -7 to +5.
- No. of analysis bands: 3 bands
- Filter Bank Range: Band 1: 80–528Hz; Band 2: 528–1941Hz; Band 3: 1941–6400Hz

- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: Sound attenuating booth; under ER-3A insert earphones, unilaterally (right ear)
- Level: Maintained at 70dBSPL with calibration
- Familiarization: Using male voice stimuli from the TIMIT database, in the corresponding conditions as used in the experiment; no feedback provided
- Testing: Each SNR in each channel for each of 120 trials (120 x 5=600 trials).
- Response mode: Repeating the sentence as accurately as possible; no feedback provided
- Outcome measure & Scoring: Percent correct identification; keywords scored as correct or incorrect if produced accurately, regardless of

order; point biserial correlation measured for all listeners; 5 keywords per sentence, hence 3000 keywords scored

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
14.	***A correlational method to concurrently measure envelope and temporal fine structure weights: Effects of age, cochlear pathology, and spectral shaping*** (Fogerty & Humes, 2012)	To measure the relative weighting given by listeners of different ages and hearing loss to E and TFS cues in continuous speech	Experimental Study	**Experiment 1** **Subjects** • Number: 8 • Age Range: 20-23 years (M=21 years) • Gender: Not specified • Language: Not specified • Ear: Monoaural • Control condition: None **Stimuli** • Speech Material: Sentences from the IEEE database, spoken by a male speaker • Conditions: 8 conditions; 2 processing conditions (E and TFS) x 4 SNRs (17, 11, 5, -1) • Noise: Speech shaped noise matched to the power spectrum of each sentence • Processing used: Hilbert transform; E and TFS were	**Results** **Experiment 1** • Listeners performed significantly better on E than TFS conditions for all SNRs • Performance plateaus at about 80% correct **Experiment 2** • E cues weighted more than TFS across bands 1 and 3; TFS cues weighted similarly across bands • Repeated sentences ○ E3 weighted more than TFS3 ○ E2 given less weight than E3

	extracted from sentences that were combined with noise to create an SNR of -5dB; for the test sentences, non-target portion was at -5dBSNR while the target portions were scaled across a range of SNRs from 5 to -7, IN 3dB steps • No. of analysis bands: 3 bands • Filter Bank Range: Band 1: 80–528Hz; Band 2: 528–1941Hz; Band 3: 1941–6400Hz • Transition of overlap: Not specified **Procedure** • Test setting & Transducer: Setting as such not specified; under ER-3A insert earphones, monoaurally • Level: Maintained at 70dBSPL with calibration • Familiarization: Using male voice stimuli from the TIMIT database, in the	○ E2 weight significantly reduced for repeated presentation **Conclusions** • E weighted more than TFS overall, however, more unstable and influenced by stimulus and cognitive factors • TFS cues more stable than E cues, although relatively less weighted

corresponding conditions as used in the experiment; no feedback provided

- Testing: 15 sentences presented for each of the 8 conditions, total of 120 sentences across conditions (600 keywords)
- Response mode: Repeating the sentence as accurately as possible; no feedback provided
- Outcome measure & Scoring: Percent correct identification; keywords scored as correct or incorrect if produced accurately, regardless of order

Experiment 2
Subjects

- Number: 8
- Age Range: 20-23 years (M=21 years)
- Gender: Not specified
- Language: Not specified

- Ear: Monoaural
- Control condition: N/A

Stimuli

- Speech Material: 600 sentences (half from Exp1, half novel, equal numbers used in the first and second half of the testing) from the IEEE database, spoken by a male speaker
- Conditions: 6 acoustic information “channels”, 3 bands x 2 processing conditions (E and TFS)
- Noise: Speech shaped noise matched to the power spectrum of each sentence
- Processing used: Hilbert transform; the SNR in each of these 6 channels were varied independently in 3dB steps from -7 to +5.
- No. of analysis bands: 3 bands

- Filter Bank Range: Band 1: 80–528Hz; Band 2: 528–1941Hz; Band 3: 1941–6400Hz
- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: Setting as such not specified; under ER-3A insert earphones, monoaurally
- Level: Maintained at 70dBSPL with calibration
- Familiarization: Using male voice stimuli from the TIMIT database, in the corresponding conditions as used in the experiment; no feedback provided
- Testing: Each SNR in each channel for each of 120 trials (120 x 5=600 trials)
- Response mode: Repeating the sentence as accurately as possible; no feedback provided

- Outcome measure & Scoring: Percent correct identification; keywords scored as correct or incorrect if produced accurately, regardless of order; point biserial correlation; 5 keywords per sentence, hence 3000 keywords scored

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
15.	***Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise.*** (Swaminathan & Heinz, 2012)	To investigate the relative contribution of E and TFS cues to speech perception in quiet and in noise	Experimental Study	**Subjects** • Number: 5 • Age Range: M=28.8 years; SD=1.9 years • Gender: Male • Language: Native speakers of American English • Ear: Monoaural; right ear • Control condition: None **Stimuli** • Speech Material: 16 /aCa/ syllables, recorded by 2 male and 2 female talkers for a total of 64 stimuli • Conditions: 5; Intact speech; Phonemic ENV (PHENV), Periodicity ENV (PDENV), Broadband TFS (BBTFS), Narrowband TFS (NBTFS) • Noise: Speech shaped noise (matching the original phoneme in amplitude spectrum)	**Results** • Across, SNRs, better scores: intact, PHENV/BBTFS, PDENV, then NBTFS • At positive SNRs, order of performance from best to worst: • Intact, PHENV, BBTFS, PDENV, to NBTFS • At negative SNRs, order of performance from best to worst: • Intact, TFS, ENV

	• Processing used: Each stimulus was noise degraded prior to vocoding. Hilbert transform; PHENV: ENV was extracted from all bands and lowpass filtered at 64Hz with 6th order Butterworth filter; PDENV: similar to PHENV, but bandpass filtered (64-300) and the first 5 bands removed; ENV conditions tone-vocoded and BP filtered and summed; BBTFS & NBTFS: TFS filtered through 1 and 16 bands respectively; the original and vocoded presented at 10, 5, 0, -5, -10, -15, -20dBSNR and in quiet (Q) • No. of analysis bands: 1 or 16 (third order Butteworth filters) • Filter Bank Range: 80-8020Hz • Transition of overlap: Not specified	• Voicing ○ Intact, PHENV/BBTFS, NBTFS/PDENV ○ Intact, TFS, ENV, at negative SNRs • Manner ○ Intact, PHENV, BBTFS, PDENV and NBTFS ○ At negative SNRs: Intact best • Place ○ At positive SNRs: Intact PHENV/BBTFS PDENV NBTFS ○ For negative SNRs: intact, BBTFS, NBTFS/PHENV, PDENV

	Procedure • Test setting & Transducer: In a double-walled sound attenuating chamber, under Sennheiser HD580 headphones monoaurally (right) • Level: 65dBSPL • Familiarization: Provided on with Q; feedback given • Testing: Each block consisted of one kind of stimulus (16 consonants x 4 voices), at a single SNR; SNR decreased progressively for subsequent blocks; test order of stimuli randomized across participants; no feedback • Response mode: Mode as such not specified • Outcome measure & Scoring: Percent correct identification; information transmission analysis	• Nasality ○ At positive SNRs: Intact BBTFS NBTFS / PHENV PDENV ○ At negative SNRs: Intact TFS PHENV PDENV **Conclusions** • E cue most important in speech perception in quiet and in noise • TFS cues also important in speech perception in noise, but likely only in combination with E cues

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
16.	***Role and relative contribution of temporal envelope and fine structure cues in sentence recognition by normal-hearing listeners.*** (Apoux et al., 2013)	To investigate the role of E and TFS cues in speech perception in noise	Experimental Study	**Subjects** - Number: 40 (2 groups) - Age Range: 19 to 26 years (Average= 21 years) - Gender: 37 female, 3 male - Language: Not specified - Ear: Diotic - Control condition: None **Stimuli** - Speech Material: 350 sentences from SPIN, half low, half high predictability - Conditions: 24 SNR combination conditions for each masker type - Noise: Masker stimuli either speech shaped noise (SSN) or AzBio sentences (SPE) - Processing used: Hilbert decomposition; with target and masker sentences mixed and chimerized at different	**Results** - For both maskers, baseline performance increases with SNR - SSN Masker: Effect of noise on the recognition of sentences due to SNRenv in -12, -6 and 0dBSNRtfs, overall - SPE Masker: Limited effect of SNRtfs **Conclusions** E cues may help in speech perception and TFS cues may help in extracting E from signal/glimpsing, in noise

SNRs (-12 to 12 (SSN), -18 to 6 (SPE) in 6dB steps)

- No. of analysis bands: 30 adjacent bands (cochlea-like)
- Filter Bank Range: 80-7563Hz (two cascaded 12th order Butterworth)
- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: Double-walled sound attenuating booth, under Sennheiser HD280 Pro circumaural headphones, diotically
- Level: 65dBA
- Familiarization: With 3 blocks of 8 sentences each, SNRs for E and TFS 1000, 6 and -6dB across blocks; feedback provided
- Testing: 24 blocks, randomized; each block consisted of 14 sentences, half low and half high predictability; 2hrs; no feedback

- Response mode: Type final word of sentence into computer
- Outcome measure & Scoring: Percent correct identification

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
17.	***Enhancement of speech intelligibility in reverberant rooms: role of amplitude envelope and temporal fine structure.*** (Srinivasan & Zahorik, 2014)	To investigate the role of E and TFS cues in speech perception in reverberant rooms, and also to ascertain if prior exposure to reverberation causes betterment of speech perception performance	Experimental Study	**Subjects** • Number: 15 (another 15 with only reverb-ENV with different PRESTO sentences, no statistical difference) • Age Range: 18.1-29.7 years (Average= 21.7 years) • Gender: Not specified • Language: Not specified • Ear: Not specified • Control condition: None **Stimuli** • Speech Material: 8 lists from PRESTO corpus, each list consisting of 18 sentences, 76 keywords in total, different talkers within each list • Conditions: 2 conditions; reverb-ENV and reverb-TFS • Noise: Reverberation; simulated rooms with BB (125-4kHz) T60: 0, 0.2, 0.3,	**Results** • Unblocked vs blocked, significant improvement only for reverb-ENV at T60 of 0.3 and 0.7s; • Reverb-TFS did not show such improvement **Conclusions** • Prior exposure to reverberation can improve speech perception scores; • E cue facilitates this speech perception rather than TFS

0.39, 0.7, 1.22, 2.32s (R0 through R6)

- Processing used: Hilbert transform; reverb-ENV: ENV extracted, convolved with different BRIR values and TFS with anechoic BRIR; vice versa with reverb-TFS; chimerized
- No. of analysis bands: 16 0.4-octave bands (6th order zero phase Butterworth filter)
- Filter Bank Range: 80-6010Hz
- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: Double-walled sound attenuating room, under Beyerdynamic DT990 Pro
- Level: 68dBA
- Familiarization: Not specified
- Testing: 288 sentences in 10 sets of trials (1-4 “blocked” with either R2 or R4, others “unblocked”, odd sets reverb-

TFS, even reverb-ENV); no feedback

- Response mode: Type out all words understood
- Outcome measure & Scoring: Percent correct identification; RAU

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
18.	***Level considerations for chimeric processing: Temporal envelope and fine structure contributions to speech intelligibility.*** (Fogerty & Entwistle, 2015)	To investigate level-related effects in chimera processing in speech perception in normal hearing young adults	Experimental Study	**Subjects** • Number: 19 (with 1 removed) • Age Range: 18-26 years • Gender: Not specified • Language: Not specified • Ear: Monoaural (right ear) • Control condition: None **Stimuli** • Speech Material: 187 SPIN sentences, in 11 blocks of 16 (half low and half high predictability) • Conditions: 11 conditions • Noise: Speech shaped noise (constant spectrum below 800Hz, 6dB/octave roll-off above • Processing used: Hilbert transform; chimera synthesizer; SNRenv combined at 6 or -6dB, SNRtfs at 6, 0 or -6dB, and	**Results** • Overall, better performance when SNRtfs and SNRenv were kept at 6dB • At 6dBSNR SNRenv, performance good regardless of SNRtfs • At -6dBSNR SNRenv, systematic difference with SNRtfs • At 0, 6dBSNR for SNRenv, SNRtfs not significantly different at 6 and -6dBSNR **Conclusions** • If E cues available, mediates speech perception with little contribution from TFS • In the event of degradation of E cues by noise, TFS can convey speech information

vice versa (reference: both at 0dBSNR)

- No. of analysis bands: 30 adjacent bands (1ERBn)
- Filter Bank Range: Not specified
- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: Sound attenuating booth, under Sennheiser HD280 Pro (right ear)
- Level: Speech maintained at 70dBA, more intensity with noise
- Familiarization: With 11 sentences across all tested conditions
- Testing: 11 randomized blocks
- Response mode: Type out final keyword of each sentence
- Outcome measure & Scoring: Percent correct identification

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
19.	***Role of Binaural Temporal Fine Structure and Envelope Cues in Cocktail-Party Listening.*** (Swaminathan et al., 2016)	To assess the relative contribution of the binaural E and TFS cues to spatial release from masking in normal hearing adults	Experimental Study	**Subjects** • Number: 10 • Age Range: 19-22 years • Gender: 9 female, 1 male • Language: Native American English speakers • Ear: (?)Binaural • Control condition: None **Stimuli** • Speech Material: Syntactically correct sentences of 5 words, spoken by 7 different talkers (female) • Conditions: Correlated TFS (Corr TFS), uncorrelated TFS (Uncorr TFS), lowpass correlated TFS (LP Corr TFS), lowpass uncorrelated TFS (LP Uncorr TFS) (LP cutoff at 1500Hz); Three spatial conditions; target and masker collocated at 0deg, or	**Results** (All dB values= mean SRM) • Corr TFS: collocated, separated scores similar for natural & 32 channel vocoded speech • Corr TFS anf LP Corr TFS: decrease in thresholds with spatial separation • Highest mean thresholds for collocated condition regardless of processing condition • Corr TFS, LP Corr TFS: steep improvement from 0deg to

	symmetrically at ±15 and ±90 - Noise: Speech on speech masking - Processing used: Hilbert transform; noise vocoding; ENV low-pass filtered below 300Hz, 4th order Butterworth, noise-vocoded, BP filtered and summed - No. of analysis bands: 8 or 32 (equal BW in log scale) - Filter Bank Range: 80-8000Hz - Transition of overlap: Not specified **Procedure** - Test setting & Transducer: In a double-walled sound attenuating chambers, under Sennheiser HD 280 headphones - Level: Maskers fixed at 55dBSPL; target level varied adaptively - Familiarization: Not specified	±15deg, less from ±15deg to ±90 deg(~4dB) ○ ~8 dB at ±15deg, ~14 dB at ±90deg - Uncorr TFS, LP Uncorr TFS: less improvement from 0 to ±15deg, more from ±15deg to ±90deg (~11dB) ○ ~2 dB at ±15deg, ~8 dB at ±90deg - Separated thresholds elevated, SRM lesser in 8 vs 32 channel condition **Conclusions** Binaural TFS cues especially in the low frequency regions (below 1500Hz) are important in mediating

	• Testing: First session: original speech material x 3 spatial configurations x 6 runs= 18 runs; then, 4 processing conditions x 2 vocoder channel conditions x 3 spatial configurations x 6 runs = 144 runs; randomized; Identify keywords from the sentence corresponding to the target (0deg) • Response mode: Mouse click on correct response • Outcome measure & Scoring: SRT, SRM; Correct if ¾ keywords identified	spatial release from masking in normal hearing individuals

Sl. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
20.	***Relative importance of temporal envelope and fine structure in lexical-tone perception (L)*** (Xu & Pfingst, 2003)	To determine whether envelope or fine structure cues are more important for lexical tone perception in Mandarin Chinese	Experimental Study	**Subjects** • Number: 5 • Age Range: 36 - 41 (37.6 ± 2.1, mean and SD) • Gender: 3 female, 2 male • Language: Native speakers of Mandarin Chinese • Ear: Binaural • Control condition: None **Stimuli** • Speech Material: 10 syllables of Mandarin Chinese and the four tone patterns (tone 1 through tone 4); these 40 CV and tone combinations formed meaningful words • Conditions: Speech-speech chimera • Noise: None • Processing used: Hilbert transform; chimaerizer; 3	**Results** • Scores for male and female talkers similar, • Across the 3 bands similar overall, yet with statistically significant differences between bands • Tone identification, on average, across 4, 8 and 16 bands: • Tone coded in fine structure: 90.8%, 89.5%, and 84.5% • Tone coded in the envelope: 4.3%, 5.0%, and 8.9% • Confusions: 4.9%, 5.5%, and 6.6% (most often between tone 3 and 4)

bands x 12 chimeric stimuli for each syllable x 10 syllables x 2 talkers (male, female) x 5 times = 3600 stimuli; test order randomized for the 3 bands

- No. of analysis bands: 4, 8 and 16 bands
- Filter Bank Range: 80-8820Hz
- Transition of overlap: 25% of the narrowest frequency band

Procedure

- Test setting & Transducer: Double-walled sound treated booth - Acoustic Systems (model RE2 242S); loudspeakers at 1m and 0 azimuth
- Level: Randomly roving between 50 and 70dBA in 5 dB steps
- Familiarization: Not specified
- Testing: Custom GUI created on MATLAB, typographical representation of the 4 tone

Conclusions

- For lexical tone perception, the fine structure is the dominant cue over the envelope, for 4-16 bands patterns and the associated Chinese characters (4AFC); loudspeaker presents one of the 12 chimeric stimuli for each syllable

- Response mode: Mouse click on response of choice
- Outcome measure & Scoring: Percentage of correct tone recognition responses

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
21.	***Relative contributions of temporal envelope and fine structure cues to lexical tone recognition in hearing-impaired listeners.*** (Wang et al., 2011)	To investigate if the relative salience for E and TFS cues to lexical tone perception was similar for individuals with normal hearing and hearing impairment, and also to draw correlations between the degree of hearing loss and the relative importance of each cue for lexical tone perception	Experimental Study	**Subjects** • Number: 22 • Age Range: 23-34 years (M=26.1; SD=2.5) • Gender: 12 female, 10 male • Language: Native Mandarin Chinese speakers • Ear: Bilaterally • Control condition: None **Stimuli** • Speech Material: A set of 16 Chinese monosyllables, with 4 tone patterns each, representing meaningful words, recorded by a male (F0: 160Hz) and a female (F0: 280Hz) Beijing Mandarin speaker • Conditions: Intact; speech-speech chimeras • Noise: None • Processing used: Hilbert transform, chimera	**Results** • Accuracy to intact tokens: 99.0% • Responses consistent with fine structure: 90.9% • With envelope: 6.8% • Not consistent with either: 2.3% • With increase in channels ○ Male voice: responses consistent with fine structure increased, less consistent with envelope ○ Female voice: responses consistent with fine structure most and that of

	synthesizer; 3 bands x 16 chimeric stimuli for each syllable x 16 syllables x 2 talkers (male, female) = 1536 stimuli; test order randomized across conditions • No. of analysis bands: 4, 8 and 16 bands • Filter Bank Range: 80-8820Hz • Transition of overlap: 25% of the narrowest filter of each band **Procedure** • Test setting & Transducer: Sound-treated booth, under headphones (model as such not specified), bilaterally • Level: 80dBSPL • Familiarization: Average of 10m, for the individual to get used to the test procedure and the apparatus • Testing: 1.5-2hrs with breaks; typographical representation of the 4 tone patterns and the associated Chinese characters (4AFC);	envelope least, for 8 channel condition • Subjects gave more responses consistent with envelope for male voice, and fine structure for female voice **Conclusions** For lexical tone perception, fine structure appears to be a more salient cue than envelope, but this relative weight depends on the spectral resolution and the fundamental frequency of the talker's voice; TFS cues may be more useful at low frequencies

token presentation randomized

- Response mode: Mouse click on response of choice
- Outcome measure & Scoring: Percentage correct tone identification

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
22.	***The Role of Temporal Envelope and Fine Structure in Mandarin Lexical Tone Perception in Auditory Neuropathy Spectrum Disorder*** (Wang et al., 2015)	To investigate the relative contributions of E and TFS cues in lexical tone perception in ANSD, and also to see if it was impaired spectral or temporal resolution that impacted TFS cues in pitch perception in ANSD	Experimental Study	**Subjects** • Number: 15 • Age Range: 23-34 years (M=26.1 years; SD= 2.5 years) • Gender: 8 female, 7 male • Language: Native speakers of Mandarin Chinese • Ear: Bilateral • Control condition: None **Stimuli** • Speech Material: 10 Chinese monosyllables, 4 tone patterns each; 40 meaningful words, spoken by an adult male (F0= 180Hz) and adult female (F0= 300Hz); 80 recorded in total • Conditions: Intact, E & TFS • Noise: None • Processing used: Hilbert transform; chimera	**Results** • Percent scores as: ○ Original token: 97.2% ○ Consistent with TFS: 92.1% ○ Consistent with E: 3.1% ○ Consistent with neither: 3.8% **Conclusions** For lexical tone perception, the fine structure is the dominant cue over the envelope (for the 16 band-condition used in this study)

synthesizer; low pass filter of 64Hz used for E generation

- No. of analysis bands: 16 bands (equally spaced on the BM)
- Filter Bank Range: 80-8820Hz
- Transition of overlap: 25% of narrowest filter

Procedure

- Test setting & Transducer: Sound treated booth, under MADSEN TDH-50P, bilaterally
- Level: 65dBSPL
- Familiarization: For 5-10 minutes
- Testing: 240 tokens (12 chimera x 10 monosyllables x 2 voices); plus 80 original tokens = 320 tokens presented in 4AFC
- Response mode: Select the tone/word they heard; ?mouse click

- Outcome measure & Scoring: Percent correct tone identification; RAU

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
23.	***Relative contributions of acoustic temporal fine structure and envelope cues for lexical tone perception in noise.*** (Qi et al., 2017)	To assess the relative contribution of E and TFS cues to lexical tone perception in noise in normal hearing listeners	Experimental Study	**Subjects** • Number: 20 • Age Range: 19-30 years (M+SD = 24.2+3.2) • Gender: 10 female, 10 male • Language: Native speakers of Mandarin • Ear: Monoaural (random ear) • Control condition: None **Stimuli** • Speech Material: 10 monosyllables, 4 tone patterns each, in a male and a female voice; duration equalized • Conditions: E and TFS; 5 SNRs (-18, -12, -6, 0, +6dB); 50 chimera (5 SNR E x 5 SNR TFS, x 2 maskers) • Noise: SSN (matched to rms of speech sample by a male and female talker), two-talker babble (TTB)	**Results** • Average tone recognition scores for SNRs of -18*, -12*, -6*, 0, +6dB: ○ SSN: 27.6%, 60.2%, 82.1%, 93.9%, and 94.7% ○ TTB: 53.5%, 72.0%, 86.4%, 92.7%, and 95.0% • For both SSN and TTB, correlation increased with SNR for both E and TFS; coefficients greater for E than for TFS at positive SNRs **Conclusions** Both E and TFS cues contribute to lexical tone perception in noise in normal hearing listeners; E cues found to contribute slightly more than TFS.

- Processing used: Hilbert transform; chimerizer; tokens mixed with tones at 5 SNRs before extracting E and TFS
- No. of analysis bands: 30 bands of 1ERB width (30 elliptical BPF)
- Filter Bank Range: 80-7563Hz
- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: Sound treated room, under Sennheiser HD280 Pro circumaural headphones
- Level: ~65dBSPL
- Familiarization: With 80 tokens from 4 SNR conditions, for the two maskers; feedback provided
- Testing: Typographical representation of the 4 tone patterns and the 4 Chinese characters; each condition: 80 tokens (80 monosyllables x 4 tones x 2 tone patterns)

- Response mode: Mouse click on option of choice
- Outcome measure & Scoring: Percent correct tone recognition; Pearson’s correlation coefficient

SI. No	Study (Author/ year)	Aim/Objective	Study Design	Method	Findings of the Study
24.	***Speech fine structure contains critical temporal cues to support speech segmentation.*** (Teng et al., 2019)	To assess if E and TFS cues represent different aspects of the speech signal; to ascertain if and how the TFS is extracted from E cues	Experimental Study	**Subjects** • Number: 21 (10 in Exp A, 11 in B, one excluded) • Age Range: A: 23-25 years; B: 22-28 years • Gender: A: 5 female, 5 male; B: 8 female, 3 male • Language: Native Chinese speakers • Ear: Not specified • Control condition: None **Stimuli** • Speech Material: 100 sentences from the Mandarin HINT, spoken by a female speaker • Conditions: 4 conditions: A: directly reversed speech (R), envelope reversed speech (ER), fine structure reversed speech (FSR), and B: envelope reversed noise-vocoded speech (ERNV) and ER	**Results** • A: Thresholds in the ER block were significantly larger than in the R block • B: Thresholds in the ER block were significantly larger than the ERNV block **Conclusions** E and TFS cues convey similar information for speech perception in quiet in normal hearing listeners

- Noise: HINT used
- Processing used: Hilbert transform; chimerizer; A; window sizes for R: 30, 50, 70, 80, 90, and 120ms; ER: 30, 70, 90, 120, 150, and 200ms; FSR: 30, 150, and 300ms; B: window sizes for ERNV: 30, 50, 70, 80, 90, and 120ms; ER: 30, 70, 90, 120, 150, and 200ms
- No. of analysis bands: 16 bands
- Filter Bank Range: 80-8820Hz
- Transition of overlap: Not specified

Procedure

- Test setting & Transducer: Soundbooth, under Sennheiser 370 headphones
- Level: ~65dBSPL
- Familiarization: Not specified
- Testing: A: 60 sentences for R, ER, 30 for FSR (x window sizes x 10 sentences), 10

sentences shared for R and ER; B: 60 ER and ERNV

- Response mode: Type out words in Excel sheet
- Outcome measure & Scoring: Phonemic scoring

All the articles included in the present review were experimental studies that included comparing the relative contribution of the E and TFS cues in their methodology, even though the main aim or research question was usually different. As can be seen in Table 3.3, the quality appraisal tool used to assess the risk of bias articles revealed similarity in methodology across the twenty-four articles selected. Researchers were never blinded to the conditions the participants were subjected to, and none provided a justification for the sample size used, which was often small. The methodology adopted by the studies are, however, heterogenous in terms of number of subjects, language of assessment, phenomenon investigated, and in the processing techniques and outcome measures used, and hence we have resorted to a qualitative description of the same. The data extracted from the final twenty-four articles are briefly outlined below.

3.1 Subject-related Parameters

The number of subjects that participated in the experiments ranged from as few as 4 to as many as 60. Most studies have divided their participant pool into subgroups and subjected each of them to different stimuli and processing conditions. Most articles have specified the age range for its participants, and atleast one of the subject groups employed fall into the category of adults. Unlike age, a far lesser number of articles have specified the gender of their participants. Whether gender is an important parameter in the perception of temporal cues in speech remains to be looked into. All studies have assessed speech perception abilities in native language of the participants. The languages were English, French or Mandarin Chinese. Monoaural presentation of speech materials to either the right ear or the ear of the participants' choice, is a rather consistent finding across studies. One article (Swaminathan et al., 2016) explicitly aimed to study binaural benefit. Of the 24 articles considered, only one has made use of a control condition in its methodology. Several articles have also mentioned the handedness of its participants, all of them being right-handed, usually by self-report. As with gender, it remains to be seen whether handedness would affect the perception of E or TFS cues.

3.2 Speech Stimuli and Processing

The speech material used in the articles comparing E and TFS cues in normal-hearing individuals were monosyllables, disyllables and sentence materials, which were for the most part meaningful. Sentences were taken from standardized lists such as the HINT or the SPIN, and the syllabic material followed the /vCv/ pattern. All articles employed a variety of speech conditions that the participants were exposed to; these would include at least one condition each for E and TFS cues. Testing in noise maskers was also carried out in many studies, where the noise was generally matched in power spectrum to the corresponding target speech material. The processing used was the Hilbert transform, which would help in chimera synthesis or vocoding. The E stimuli thus generated would be modulating either a noise or a sine wave carrier. The TFS stimuli was often power-weighted. Other novel processing schemes were also used along with vocoding or chimera synthesis, such as envelope expansion/compression, etc. Although the number of analysis bands for the studies varied over a wide range, from 1 to 64, across the studies considered, 4-16 was a number used in most of the studies. Octave, half-octave, 0.35-octave or quarter-octave-wide bands were used. The filter bank range was also mostly consistent with 80-8820Hz in most articles, although exceptions existed. The transition of overlap, wherever specified, was generally 25% of the narrowest frequency band of the filters. Almost all the articles specified the usage of the Greenwood map to decide the distribution of the center frequencies of the filterbanks used.

3.3 Procedural Variables

The testing was always carried out in a sound-proof booth (with the exception being one study, where a "quiet room" was used). Headphones were the transducer of choice, and presentation, as was mentioned earlier, was mostly monoaural. Presentation was at a comfortable level. Most of the studies familiarized the participants with the stimuli and the procedure. The test procedures utilized variable number of stimuli, processing conditions, filter bands and signal-to-noise ratios (if the testing involved the use of noise). Randomization of these parameters was carried out in many studies, with two articles specifying the use of a Williams square and a Latin square, respectively, for the randomization procedure. Some articles have mentioned that the perceptual difficulty of the processing conditions move from easy to hard, while others have used the opposite order, citing avoidance of learning effects. The E and TFS conditions in many studies were counterbalanced across subjects. The mode of response for studies were usually mouse click on the desired response, repeating the heard material verbatim, or occasionally, typing out the entire sentence heard. Outcome measures generally included percent correct recognition of the stimuli, speech reception threshold, and/or information transmission analyses.

3.4 Results Across Studies

There is a general consensus among studies that, for speech perception in quiet, the E cues dominate over the TFS. This is true regardless of the language of testing, so long as the language is a non-tonal one. This dynamic changes when noise is present. In presence of noise TFS cues become important for speech perception, but this seems to be true only if E cues are also present. In lexical tone perception TFS cues are more important relative to E cues in quiet condition (Smith et al., 2002). In the presence of noise E cues help mediate lexical tone perception. These conclusions are gross at best, and a finer assessment of the relative importance of E and TFS cues in different conditions will be carried out in the following section.

Table 3.3

Results from the Risk of Bias Assessment Tool for the Articles included in the Study

Questions	**Drullman (1995)**	**Xu and Pfingst (2003)**	**Lorenzi et al. (2006)**	**Gilbert et al. (2007)**	**Sheft et al. (2008)**	**Bertoncini et al. (2009)**	**Lorenzi et al. (2009)**	**Ardoint and Lorenzi (2010)**
Q1. Was the aim/objective of the study clearly defined?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Q2. Were the participant inclusion criteria clearly described?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Q3. Are the main study findings explained?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Q4. Are the main outcome measures clearly stated?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Q5. Were the investigators blinded to the participant characteristics to reduce bias?	No	No	No	No	No	No	No	No
Q6. Is there a clarification for the appropriateness of the sample size studied?	No	No	No	No	No	No	No	No

Q7. Have the investigators provided a clarification about the settings under which the findings can be applied?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Questions	**Gilbert and Lorenzi (2010)**	**Eaves et al. (2011)**	**Fogerty (2011)**	**Wang et al. (2011)**	**Fogerty and Humes (2012a)**	**Fogerty and Humes (2012b)**	**Swaminathan and Heinz (2012)**	**Apoux et al. (2013)**
Q1. Was the aim/objective of the study clearly defined?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Q2. Were the participant inclusion criteria clearly described?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Q3. Are the main study findings explained?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Q4. Are the main outcome measures clearly stated?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Q5. Were the investigators blinded to the participant characteristics to reduce bias?	No	No	No	No	No	No	No	No
Q6. Is there a clarification for the appropriateness of the sample size studied?	No	No	No	No	No	No	No	No
Q7. Have the investigators provided a clarification about the settings under which the findings can be applied?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Questions	**Srinivasan and Zahorik, (2014)**	**Fogerty and Entwistle (2015)**	**Wang et al. (2015)**	**Swaminathan et al. (2016)**	**Qi et al. (2017)**	**Wirtzfeld et al. (2017)**	**Hou & Xiu (2018)**	**Teng et al. (2019)**
Q1. Was the aim/objective of the study clearly defined?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Q2. Were the participant inclusion criteria clearly described?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Q3. Are the main study findings explained?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Q4. Are the main outcome measures clearly stated?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Q5. Were the investigators blinded to the participant characteristics to reduce bias?	No	No	No	No	No	No	No	No
Q6. Is there a clarification for the appropriateness of the sample size studied?	No	No	No	No	No	No	No	No
Q7. Have the investigators provided a clarification about the settings under which the findings can be applied?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Chapter 4

Discussion

The present review aimed to study the relative importance of the E and TFS cues to speech and lexical tone perception in quiet and adverse listening conditions. We will discuss the results of the studies under three separate headings.

4.1 Speech Perception in Quiet Listening Conditions

The articles that look at speech perception in quiet have consensus regarding the role of E and TFS cues in speech perception in quiet. Firstly, all studies unequivocally support the notion that speech perception with intact stimuli is superior to that with either E or TFS speech alone, showing that perfect intelligibility requires both E and TFS cues (Gilbert et al., 2007; Lorenzi et al., 2006). Secondly, when processed with a fairly large number of bands (16 in most cases), E cues are necessary and sufficient to transmit speech information in quiet (Bertoncini et al., 2009; Sheft et al., 2008). TFS cues can also independently result in moderate to high levels of consonant identification when similarly processed (Lorenzi et al., 2006).

The number of analysis bands used is one factor that affects the perception of E and TFS speech. Generally, the perception of E speech improves with increasing number of frequency bands and plateaus at around 6-16 bands (Smith et al., 2002; Wirtzfeld et al., 2017), although moderate levels of consonant perception is possible with just 3 bands (Shannon et al., 1995). As noted in the investigations of Smith et al. (2002), for TFS speech, perception scores decrease with increase in number of analysis bands. One of the reasons for this may be due to the fact that, for speech processed with one or two wide bands, the envelope may be recovered from the peripheral (cochlear) auditory filters (Ghitza, 2001). Most articles we have considered have, however, controlled for this phenomenon by using a larger number of analysis bands, and some have conducted additional experiments to find out if the improvement demonstrated in speech perception with their TFS speech was due to envelope recovery. This has been carried out by passing the TFS speech through a set of gammatone (Patterson et al., 1987) or gammachirp filters (Irino & Patterson, 1997) to arrive at the recovered envelope, and testing the intelligibility of this recovered envelope in normal hearing listeners. In all such cases, the intelligibility of such stimuli was too low to account for the intelligibility of TFS speech, and it has been concluded that TFS cues in and of itself can provide speech information (Sheft et al., 2008).

Training is an important factor when it comes to understanding such processed speech, since the human auditory system is not used to extracting information from speech that contains only TFS cues (Moore, 2008). Familiarization with the stimuli is required less for E than for TFS speech to reach comparable levels of performance; on average, E speech seems to require 4-10 sessions, and more than 10 sessions are required for TFS speech (Lorenzi et al., 2006; Sheft et al., 2008). It may be speculated that the spectral resolution abilities of the individual may be a determining factor when it comes to TFS speech perception, since the auditory system tracks the fine spectral changes occurring in time to process this frequency modulated stimulus (Teng et al., 2019). It has also been noted that separate training need not be provided; instead, if the actual test is of a long duration, such as one hour, the familiarization process may take place within the first half hour (Wirtzfeld et al., 2017). However, even with significant training, the scores with E cues alone are generally greater than that with only TFS cues. A carrier phrase, if used during stimulus presentation also act as a primer for the auditory system and may facilitate familiarization with either processing (Wirtzfeld et al., 2017).

Different categories of phonetic information may be conveyed by E or TFS cues, as noted by Rosen (1992), who said that manner is better represented by E and place by

TFS cues. This is consistent with information transmission analyses data from experiments, which revealed that in quiet, E cues are capable of conveying nasality and voicing cues, and to a lesser extent manner, and then place information. For TFS speech, nasality and voicing are conveyed as is the case with E cues, however, place is conveyed better than manner (Bertoncini et al., 2009). This may also be dependent on the language, since in French, the language of testing in the above case, all nasals are voiced, and hence manner could be cued by voicing cues as well.

Although most studies are on response accuracy, response latency has also been looked at, and it has been seen that E speech is identified quicker than TFS speech. Indeed, in one study, it has been noted by participants that the TFS speech stimuli used did not sound like speech at all (Teng et al., 2019). This correlates well with the fact that TFS cues need a longer familiarization period.

The characteristics of the noise used in chimera processing may also affect speech perception based on whether it is present in the E or the TFS part of the chimera. If a noise that is made to match the rms of the signal is used, as is the case with most studies using noise in chimera synthesis, the presence of the noise in the TFS does not degrade intelligibility as much as would have been the case were it used as the E portion. The use of such noise as the TFS seems to be facilitating intelligibility presumably because of the correlation between this and the corresponding E. However, when such a noise is used in the E, there may be a resultant strong spectral tilt that appears to disrupt the speech perception with this stimulus (Wirtzfeld et al., 2017). It has thus been suggested that a wideband noise be used instead of such matched noise to avoid such influences in chimera speech production. Such a suggestion is at odds with the processing that has been used in the classic study of (Smith et al., 2002), who found that noise as the TFS is less disruptive than a speech signal. This particular finding may have been in part due to the facilitative nature of the matched noise TFS used in their study.

For TFS processing, restriction of the instantaneous frequency excursions of the TFS affects speech perception more than manipulations such as randomization of starting phase of the TFS. Although it is these frequency excursions that code envelope (as implied by the investigations of Gilbert and Lorenzi (2006)), reduction of E recovery with such band-limiting may not be responsible for this degradation in speech scores. Such frequency variations are an inherent characteristic of the TFS speech stimulus that codes for different speech features, and any modifications herein may influence the perception of speech since the acoustic characteristics cannot be perceived accurately (Sheft et al., 2008).

Most speech recognition experiments used identification of consonants and vowels in a /VCV/ context. Studies have shown that vowels are the carriers of intelligibility in the sentence context (Fogerty & Kewley-Port, 2009; Kewley-Port et al., 2007). The temporal cue that determines the relative weighting of vowels in the sentence context was found not to be limited to just the fundamental frequency contour of the sentence, but also to the temporal parameters of TFS and E cues. Specifically, in the context of sentences, E cues contribute to both vowel and consonant identification with vowels significantly more than consonants, but TFS cues do the same only with vowels. For isolated words, both vowel and consonant information are similarly conveyed by E cues, but as is the case with sentences, the TFS carries only vowel information. Across contexts, significantly more information is carried for vowels in sentences than vowels in words, and such a difference is not seen with consonants. This relative importance of vowel information in sentences may be mediated by E cues (Fogerty & Humes, 2012a). Specifically, it has been said that E cues are dominant at around 4Hz, which is also the syllabic rate of English, and thus E cues may aid in syllabification (Rosen, 1992). TFS cues seem to exist exclusively for vowels and not consonants in both word and sentence contexts (Fogerty & Humes, 2012a); the reason for this remains unexplored. Additionally, it has been seen that with number of bands less than six, E speech facilitated consonant perception more than vowels, while the reverse was generally true for TFS speech (Wirtzfeld et al., 2017).

It has been noted that the temporal cues are fairly robust to extraneous manipulations, but the degree of stability is dependent on the specific condition. One such parameter may be the level of the stimulus. Most studies have used a comfortable presentation level of around 65-80dBA. However, one study has used a level of 45dBSPL to investigate changes that would occur with this condition. Lower listening levels are associated with higher frequency resolution (Glasberg & Moore, 2000; Rhode, 1971; Robles et al., 1986), and this may assist the recovery of envelope information at this level. It has been seen, however, that not only is there no much E recovery, but also, the identification performance is also not significantly affected for the E or any of the TFS conditions (Sheft et al., 2008). Hence the E and TFS information are also relatively robust to changes in stimulus level (Eaves et al., 2011).

Although the mean identification performance did not reduce with a decrease in presentation level, the reception of specific features appears to be significantly affected. Of note, manner perception with E cues and place perception with TFS processed with starting phase randomization (the PMz condition) was significantly reduced with decrease in level to 45dbSPL (Sheft et al., 2008).

The counterbalancing that may be used in experiments makes effects of order of testing an important parameter to consider. TFS speech perception seems more robust to order effects than E speech. With E condition first, the identification scores for E speech are significantly better, as opposed to when TFS is presented initially. However, no significant effect is noted for TFS speech in either order. Order significantly affects latency as well; again, the effect is only seen for E than TFS speech. When E speech is presented first, E identification responses seem to be more delayed than with the initial presentation of TFS speech. The latency of response for TFS speech is seen not to vary with order, and appears to be at a constant level of around 400ms (Bertoncini et al., 2009). On the whole, TFS cues, even though seems to play a relatively minor role as compared to E cues, appears to be more resistant to extraneous variables.

Segmental duration is also an important consideration in the perception of E and TFS speech. With the presence of only TFS cues and no much amplitude information, intelligibility of TFS speech seems to rely on the recovered E cues, since performance improves with wider analysis bands (4 to 2 to 1 ERB). With the amplitude information being equal to the rms of the E, as is used in most studies involving TFS speech, along with E recovery, segmental duration also becomes important. Speech perception is high when amplitude fluctuations of 20Hz is present, regardless of spectral information (number of bands). When predominantly E cues are present, speech intelligibility drops precipitously as amplitude information decreases below 10Hz. Hence amplitude information above 10Hz is important for mediating speech perception with E cues, and there appears to be a trade-off between spectral and temporal information present in speech perception (Hou & Xu, 2018).

4.2 Speech Perception in Adverse Listening Conditions

While speech perception in quiet can be exclusively maintained by E cues, speech perception in noise appears to be mediated predominantly by TFS cues. This is the result of studies that show that speech perception scores in noise decreases appreciably when replacing the TFS of a speech material by noise, such that only the E remains. This effect is more pronounced when the background is a fluctuating one (Gnansia et al., 2009; Nelson et al., 2003; Stickney et al., 2005), and has resulted in the characterization of the TFS as especially important in mediating speech perception during speech-on-speech masking.

The contention that TFS is the primary mediator of intelligibility in noise has been opposed by studies that assert that such an effect is due to the specific processing procedures involved. It has been shown by Apoux and colleagues that vocoder processing as used in the studies above, may have led to spurious artefacts. Firstly, the use of a single carrier is misleading since in this case, two different envelopes now modulate a single carrier. Such absence of one of the carriers prohibits stream segregation, which would result in a “sorting problem” for the listener. Another issue associated with vocoder processing is the use of the random noise carrier. This carrier is not in reality the TFS of either the speech or the masker signal, and would result in several inconsistencies. The fact that the TFS is related to neither target or masker envelope would cause a mismatch in the information in the TFS and the envelope (Apoux et al., 2013; Apoux & Healy, 2011, 2013). This particular issue may be supported by the observation made by Smith et al. (2002) that the perception of speech-noise chimeras is easier than that of speech—speech chimeras, meaning to say that the presence of information in the TFS can disrupt intelligibility of the E, presumably due to the recovered envelope. Since the speech in the E part of the speech-speech chimeras is identified more often than that in the TFS, true envelopes may be considered to contribute more to intelligibility than recovered E (Apoux et al., 2013).

As mentioned before, a solution to this problem is to use more than one carrier (Apoux & Healy, 2013). The procedure used by Smith et al. (2002) is modified in that both the E and TFS of the chimera is now just the same speech-noise mixture, but combined at different SNRs in both the features. Results of studies using such a processing has shown interesting findings. One, the E cue is the primary cue that facilitates speech perception, even at unfavourable SNRs. The TFS cue becomes important only at low SNR levels of the E, and this effect, though significant, is small (Fogerty & Entwistle, 2015). Second, preserving only the target TFS alone is just as detrimental to speech intelligibility as keeping just the masker TFS (Apoux et al., 2013). This finding goes to say that TFS cues by itself does not mediate intelligibility, but instead helps the auditory system track the individual stream in a mixture of sounds (Apoux & Healy, 2013; Teng et al., 2019). As noted in the vocoder processing studies, the effect of the TFS may be more substantial in fluctuating maskers (as seen in Eaves et al. (2011) as well) since these cues allow for listening in the dips of the masker, in line with the "glimpsing" model proposed by Cooke (2006).

It has also been noted that preserving the TFS of the target improves intelligibility by 1dB compared to adding noise to both the E and TFS. When an artificial noise floor envelope is used and TFS preserved, there is improvement of another 5-6dB. This shows that the intelligibility enhancement provided by just the TFS is minimal, and noise degrades the envelope more than reduction of modulations therein (Drullman, 1995).

Information transmission analysis data is slightly different for speech perception in noise versus quiet. In quiet, it is seen that the E cues are capable of conveying manner, nasality, voicing and place, but these are taken over by TFS cues when noise is present. Manner reception is extremely low at adverse SNRs (Swaminathan & Heinz, 2012).

Studies have shown that the use of missing speech segments in the stimuli is less detrimental to the overall intelligibility of the sentence than filing those regions with E cues provided by noise vocoded speech. Additionally, it has been noted that abolishing TFS cues as in the noise vocoded segments of such sentences affects intelligibility by 20 percentage points when compared to when both E and TFS cues are retained as in the unprocessed condition. E cues are therefore unable to fill in the gaps left by missing speech information and instead prove to have a disruptive effect to speech understanding, and the use of additional TFS information is required along with E cues to provide high levels of speech understanding (Gilbert & Lorenzi, 2010). It should be noted, however, that this study was carried out using vocoder processing, which has the limitations stated above. Specifically, the vocoded material used may have confounded the results due to the interaction of the random noise carrier with the speech envelope.

Like noise, reverberation also impedes speech transmission. It has been evidenced that prior exposure to the reverberant environment can improve intelligibility. Even though it is the TFS cue that has been shown to be involved in stream segregation, this improvement in intelligibility on subsequent exposure to the adverse listening condition of reverberation is brought about by the E cues, and this improvement occurs rapidly, over the order of several seconds (Srinivasan & Zahorik, 2014). This further shows that E cues are plastic by nature and can be modified by various top-down factors, unlike the more stable TFS cues.

When only the E or TFS information lesser than 1.5kHz are retained, the intelligibility of TFS cues suffer more than E cues (Lorenzi et al., 2009). This may be because TFS cues have more information in the low frequency range as compared to the E cues, owing to the limitations of neural phase-locking in mammals (Johnson, 1980; Kiang et al., 1965). Perhaps related to this, another interesting phenomenon that can be noted with the TFS is that the original carrier need not be preserved if intelligibility is to be maintained in conditions where the masker and target are spatially separated. Instead, it is enough that correlated TFS stimuli reach both the ears, and it is important that this correlation is at the lower frequencies, i.e. below 1.5kHz (Swaminathan et al., 2016).

The importance of different frequency bands to the intelligibility of E and TFS cues has also been investigated. It has been seen that overall, listeners provide more weight to the E than TFS cues in perception, and more at the midfrequency region than low or high frequencies (Fogerty, 2011; Fogerty & Humes, 2012b). It has also been noted that while the perceptual crossover frequency is similar and around 1.5kHz for both E and TFS cues, the gradients are significantly greater than 0 only for E cues (Ardoint & Lorenzi, 2010). This means that the most perceptually important frequency region for TFS speech is from around 1-2.5kHz. However, the relative weights of different frequency bands seem to be affected by top-down factors such as cognition and familiarity. When repeated sentences are used, the relative weights across frequency for the E cues decreases for the midfrequency and shifts more to the lower and higher frequencies. The relative weights for the TFS are comparatively flatter across the spectrum and more resistant to such extraneous modifications (Fogerty & Humes, 2012b). TFS cues therefore seem more robust to perturbations than E cues. E cues may also be targeted in auditory training programs due to this malleability, as noted by the authors of the study.

Periodic interruptions imposed on the E and TFS shows that both these cues are quite robust to such modifications. The effect of such interruptions is more for a square than a sine wave, and is more for E than TFS cues, especially at lower modulation frequencies (<16Hz) (Gilbert et al., 2007). As stated before, TFS cues appear to be more stable to extraneous manipulations as compared to the E cues.

4.3 Lexical Tone Perception in Quiet and in Adverse Listening Conditions

Although part of speech perception, lexical tone perception as an auditory task is more similar to melody recognition. Smith et al. (2002) has noted that for melody recognition, TFS cues are the more important contributor. Hence, unlike the case for speech perception, lexical tone perception is likely to be predominantly mediated by TFS cues in quiet, and there has been empirical evidence for the same. When a speech-speech chimera involving Mandarin tones is presented, listeners tend to hear the tone represented in the TFS part (Wang et al., 2015). This result is generally similar over 4-16 bands. With increase in band number, the importance of the E portion also increases in relative salience (Wang et al., 2011; Xu & Pfingst, 2003).

Some studies have reported that the speaker may also influence which cue is utilized predominantly. Female speakers have a higher fundamental frequency than males, and since higher frequencies are resolved more than lower ones in the normal auditory filters, female voice seems to be more dominated by TFS cues, and male by E cues (Wang et al., 2011).

For lexical tone perception in noise, it is seen that both the E and the TFS cues contribute, with the relative weight of E cues being slightly more than TFS (Qi et al., 2017).

4.4 Limitations of the Study

The present review has aimed to study the relative contribution of E and TFS cues to speech and lexical tone perception. There are several limitations that may have affected the complete realization of this aim:

- The keywords employed in the present review has not covered the alternative nomenclatures that may be used by several articles for the E and TFS cues such as “ENV” or “FS” or “amplitude modulations”, “AM”, “frequency modulations”, “FM”, etc.
- “Lexical tone perception” as a keyword has not been employed

- Periodicity cues may have significantly confounded the results of the present study, since the Hilbert E and TFS used in most articles studying the phenomenon incorporate this temporal cue within its frequency range.

Chapter 5

Summary and Conclusion

The present review aimed to study the relative importance of E and TFS cues in speech and lexical tone perception across quiet and adverse listening conditions. A systematic review was conducted wherein, following screening and exclusion from the initial pool of 9,138, there resulted 24 records relevant to the research question at hand. These 24 articles were assessed for methodological rigour, and it was found that all articles had similar limitations with respect to experimental procedure and thus appear to be at equivalent level of evidence. A qualitative synthesis of information from these articles revealed that E and TFS cues play different roles in different conditions. For speech perception in quiet, E, and to relatively smaller extent, TFS cues, can mediate intelligibility, while it is seen that E is the primary determinant of the same when noise is present. TFS cues are helpful in modulated maskers such as speech babble because it provides a way for the auditory system to listen in the dips of the masker and thus extract E information from the same. E and TFS cues also convey phonetically different information, and although not the dominant carrier of intelligibility, TFS cues are generally more robust than E cues to extraneous stimulus manipulations. For lexical tone perception, TFS mediates tone identification in quiet, while in noise, the same is carried out by both E and TFS cues with slightly more contribution from E. On the whole, neither E nor TFS cues can exclusively maintain speech perception abilities in all conditions, and both these temporal cues contribute to the robustness of speech perception seen across the spectrum of listening conditions.

Present day multichannel cochlear implants have reduced spectral resolution (Oxenham & Kreft, 2014), and work predominantly on temporal cues. Progress in identifying and maximizing these cues may have important implications for the future of signal-processing schemes used in cochlear implants.

5.1 Implications

- May have direct implications for speech-processing strategies in cochlear implants
- May assist in patient and caregiver counselling
- A better understanding of the speech perception mechanism in the healthy human auditory system may indirectly assist in furthering the progress made in the field of speech recognition algorithms and software

References

Apoux, F., & Healy, E. W. (2011). Relative contribution of target and masker temporal fine structure to the unmasking of consonants in noise. *The Journal of the Acoustical Society of America*, *130*(6), 4044–4052.

Apoux, F., & Healy, E. W. (2013). A glimpsing account of the role of temporal fine structure information in speech recognition. *Advances in Experimental Medicine and Biology*, *787*, 119–126. https://doi.org/10.1007/978-1-4614-1590-9_14

Apoux, F., Yoho, S. E., Youngdahl, C. L., & Healy, E. W. (2013). Role and relative contribution of temporal envelope and fine structure cues in sentence recognition by normal-hearing listeners. *The Journal of the Acoustical Society of America*, *134*(3), 2205–2212. https://doi.org/10.1121/1.4816413

Ardoint, M., & Lorenzi, C. (2010). Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine structure or envelope cues. *Hearing Research*, *260*(1), 89–95. https://doi.org/10.1016/j.heares.2009.12.002

Bertoncini, J., Serniclaes, W., & Lorenzi, C. (2009). Discrimination of speech sounds based upon temporal envelope versus fine structure cues in 5- to 7-year-old children. *Journal of Speech, Language, and Hearing Research : JSLHR*, *52*(3), 682–695. https://doi.org/10.1044/1092-4388(2008/07-0273)

BSA. (2004). *British Society of Audiology Recommended Procedure, Pure Tone Air and Bone Conduction threshold, Audiometry with and without Masking, and determination of Uncomfortable Loudness Levels*.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573.

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology & Community Health*, *52*(6), 377–384.

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *The Journal of the Acoustical Society of America*, *97*(1), 585–592. https://doi.org/10.1121/1.413112

Eaves, J. M., Quentin Summerfield, A., & Kitterick, P. T. (2011). Benefit of temporal fine structure to speech perception in noise measured with controlled temporal envelopes. *The Journal of the Acoustical Society of America*, *130*(1), 501–507. https://doi.org/10.1121/1.3592237

Faulkner, A., Rosen, S., & Smith, C. (2000). Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, *108*(4), 1877–1887.

Fogerty, D. (2011). Perceptual weighting of individual and concurrent cues for sentence intelligibility: Frequency, envelope, and fine structure. *The Journal of the*

Acoustical Society of America, *129*(2), 977–988. https://doi.org/10.1121/1.3531954

Fogerty, D., & Entwistle, J. L. (2015). Level considerations for chimeric processing: Temporal envelope and fine structure contributions to speech intelligibility. *The Journal of the Acoustical Society of America, 138*(5), EL459-464. https://doi.org/10.1121/1.4935079

Fogerty, D., & Humes, L. E. (2012a). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, *131*(2), 1490–1501. https://doi.org/10.1121/1.3676696

Fogerty, D., & Humes, L. E. (2012b). A correlational method to concurrently measure envelope and temporal fine structure weights: Effects of age, cochlear pathology, and spectral shaping. *The Journal of the Acoustical Society of America*, *132*(3), 1679–1689. https://doi.org/10.1121/1.4742716

Fogerty, D., & Kewley-Port, D. (2009). Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *The Journal of the Acoustical Society of America*, *126*(2), 847–857.

Ghitza, O. (2001). On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. *The Journal of the Acoustical Society of America*, *110*(3), 1628–1640.

Gilbert, G., Bergeras, I., Voillery, D., & Lorenzi, C. (2007). Effects of periodic interruptions on the intelligibility of speech based on temporal fine-structure or envelope cues. *The Journal of the Acoustical Society of America*, *122*(3), 1336. https://doi.org/10.1121/1.2756161

Gilbert, G., & Lorenzi, C. (2006). The ability of listeners to use recovered envelope cues from speech fine structure. *The Journal of the Acoustical Society of America*, *119*(4), 2438–2444.

Gilbert, G., & Lorenzi, C. (2010). Role of spectral and temporal cues in restoring missing speech information. *The Journal of the Acoustical Society of America*, *128*(5), EL294-299. https://doi.org/10.1121/1.3501962

Glasberg, B. R., & Moore, B. C. (2000). Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise. *The Journal of the Acoustical Society of America*, *108*(5), 2318–2328.

Gnansia, D., Péan, V., Meyer, B., & Lorenzi, C. (2009). Effects of spectral smearing and temporal fine structure degradation on speech masking release. *The Journal of the Acoustical Society of America*, *125*(6), 4023–4033.

Gunjawate, D. R., Ravi, R., & Bellur, R. (2018). Acoustic analysis of voice in singers: A systematic review. *Journal of Speech, Language, and Hearing Research, 61*(1), 40–51.

Hilbert, D. (1912). *Grundzuge einer allgemeinen Theorie der linearen Integralgleichungen*.

Hou, L., & Xu, L. (2018). Role of short-time acoustic temporal fine structure cues in sentence recognition for normal-hearing listeners. *The Journal of the Acoustical Society of America, 143*(2), EL127. https://doi.org/10.1121/1.5024817

Irino, T., & Patterson, R. D. (1997). A time-domain, level-dependent auditory filter: The gammachirp. *The Journal of the Acoustical Society of America*, *101*(1), 412–419.

Johnson, D. H. (1980). The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *The Journal of the Acoustical Society of America*, *68*(4), 1115–1122.

Kewley-Port, D., Burkle, T. Z., & Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *122*(4), 2365–2375.

Kiang, N. Y., Pfeiffer, R. R., Warr, W. B., & Backus, A. S. (1965). *XLI Stimulus Coding in the Cochlear Nucleus*. SAGE Publications Sage CA: Los Angeles, CA.

Kong, Y.-Y., & Zeng, F.-G. (2006). Temporal and spectral cues in Mandarin tone recognition. *The Journal of the Acoustical Society of America, 120*(5 Pt 1), 2830–2840. https://doi.org/10.1121/1.2346009

Lorenzi, C., Debruille, L., Garnier, S., Fleuriot, P., & Moore, B. C. J. (2009). Abnormal processing of temporal fine structure in speech for frequencies where absolute thresholds are normal. *The Journal of the Acoustical Society of America*, *125*(1), 27–30. https://doi.org/10.1121/1.2939125

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. J. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences, 103*(49), 18866–18869. https://doi.org/10.1073/pnas.0607364103

Lorenzi, C., & Moore, B. C. (2007). Role of temporal envelope and fine structure cues in speech perception: A review. *Proceedings of the International Symposium on Auditory and Audiological Research*, *1*, 263–272.

Moore, B. C. J. (2008). The Role of Temporal Fine Structure Processing in Pitch Perception, Masking, and Speech Perception for Normal-Hearing and Hearing-Impaired People. *Journal of the Association for Research in Otolaryngology*, *9*(4), 399–406. https://doi.org/10.1007/s10162-008-0143-x

Nelson, P. B., Jin, S.-H., Carney, A. E., & Nelson, D. A. (2003). Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America*, *113*(2), 961–968.

Oxenham, A. J., & Kreft, H. A. (2014). Speech perception in tones and noise via cochlear implants reveals influence of spectral resolution on temporal processing. *Trends in Hearing*, *18*, 2331216514553783.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., & Brennan, S. E. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Bmj*, *372*.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., & Rice, P. (1987). An efficient auditory filterbank based on the gammatone function. *A Meeting of the IOC Speech Group on Auditory Modelling at RSRE*, *2*(7).

Qi, B., Mao, Y., Liu, J., Liu, B., & Xu, L. (2017). Relative contributions of acoustic temporal fine structure and envelope cues for lexical tone perception in noise. *The Journal of the Acoustical Society of America*, *141*(5), 3022. https://doi.org/10.1121/1.4982247

Rhode, W. S. (1971). Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique. *The Journal of the Acoustical Society of America*, *49*(4B), 1218–1231.

Robles, L., Ruggero, M. A., & Rich, N. C. (1986). Basilar membrane mechanics at the base of the chinchilla cochlea. I. Input–output functions, tuning curves, and response phases. *The Journal of the Acoustical Society of America*, *80*(5), 1364–1374.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *336*(1278), 367–373.

Sanderson, S., Tatt, I. D., & Higgins, J. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. *International Journal of Epidemiology*, *36*(3), 666–676.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Sheft, S., Ardoint, M., & Lorenzi, C. (2008). Speech identification based on temporal fine structure cues. *The Journal of the Acoustical Society of America*, *124*(1), 562–575. https://doi.org/10.1121/1.2918540

Shetty, H. N. (2016). Temporal cues and the effect of their enhancement on speech perception in older adults–A scoping review. *Journal of Otology*, *11*(3), 95–101.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87–90. https://doi.org/10.1038/416087a

Srinivasan, N. K., & Zahorik, P. (2014). Enhancement of speech intelligibility in reverberant rooms: Role of amplitude envelope and temporal fine structure. *The Journal of the Acoustical Society of America*, *135*(6), EL239-245. https://doi.org/10.1121/1.4874136

Stickney, G. S., Nie, K., & Zeng, F.-G. (2005). Contribution of frequency modulation to speech recognition in noise. *The Journal of the Acoustical Society of America, 118*(4), 2412–2420.

Swaminathan, J., & Heinz, M. G. (2012). Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(5), 1747–1756. https://doi.org/10.1523/JNEUROSCI.4493-11.2012

Swaminathan, J., Mason, C. R., Streeter, T. M., Best, V., Roverud, E., & Kidd, G. J. (2016). Role of Binaural Temporal Fine Structure and Envelope Cues in Cocktail-Party Listening. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *36*(31), 8250–8257. https://doi.org/10.1523/JNEUROSCI.4421-15.2016

Teng, X., Cogan, G. B., & Poeppel, D. (2019). Speech fine structure contains critical temporal cues to support speech segmentation. *NeuroImage*, *202*, 116152. https://doi.org/10.1016/j.neuroimage.2019.116152

Tyler, R. S. (1988). Open-Set Word Recognition With the 3M/Vienna Single-Channel Cochlear Implant. *Archives of Otolaryngology–Head & Neck Surgery*, *114*(10), 1123–1126. https://doi.org/10.1001/archotol.1988.01860220057023

Wang, S., Dong, R., Liu, D., Wang, Y., Liu, B., Zhang, L., & Xu, L. (2015). The Role of Temporal Envelope and Fine Structure in Mandarin Lexical Tone Perception in Auditory Neuropathy Spectrum Disorder. *PLOS ONE*, *10*(6), e0129710. https://doi.org/10.1371/journal.pone.0129710

Wang, S., Xu, L., & Mannell, R. (2011). Relative contributions of temporal envelope and fine structure cues to lexical tone recognition in hearing-impaired listeners. *Journal of the Association for Research in Otolaryngology : JARO*, *12*(6), 783–794. https://doi.org/10.1007/s10162-011-0285-0

Wirtzfeld, M. R., Ibrahim, R. A., & Bruce, I. C. (2017). Predictions of Speech Chimaera Intelligibility Using Auditory Nerve Mean-Rate and Spike-Timing Neural Cues. *Journal of the Association for Research in Otolaryngology : JARO*, *18*(5), 687–710. https://doi.org/10.1007/s10162-017-0627-7

Xu, L., & Pfingst, B. E. (2003). Relative importance of temporal envelope and fine structure in lexical-tone perception. *The Journal of the Acoustical Society of America*, *114*(6 Pt 1), 3024–3027. https://doi.org/10.1121/1.1623786

Zeng, F.-G., Nie, K.-B., Stickney, G., Liu, S., Rio, E. D., Kong, Y.-Y., & Chen, H.-B. (2003). Facts and artifacts in auditory chimaeras. *Assoc. Res. Otolaryngol. Abstr*, *26*, 213.