Role of short-time acoustic temporal fine structure cues in sentence recognition for normal-hearing listeners

Limin Hou
Communication and Information Engineering, Shanghai University, Shanghai, China
lmhou@staff.shu.edu.cn

Li Xu
Communication Sciences and Disorders, Ohio University, Athens, Ohio 45701, USA
xul@ohio.edu

Abstract: Short-time processing was employed to manipulate the amplitude, bandwidth, and temporal fine structure (TFS) in sentences. Fifty-two native-English-speaking, normal-hearing listeners participated in four sentence-recognition experiments. Results showed that recovered envelope (E) played an important role in speech recognition when the bandwidth was $>1$ equivalent rectangular bandwidth. Removing TFS drastically reduced sentence recognition. Preserving TFS greatly improved sentence recognition when amplitude information was available at a rate $/C21^{/C20}$10 Hz (i.e., time segment $/C20^{/C20}$100 ms). Therefore, the short-time TFS facilitates speech perception together with the recovered E and works with the coarse amplitude cues to provide useful information for speech recognition.

1. Introduction

Temporal information in an acoustic signal can be partitioned into two components, i.e., the rapidly varying temporal fine structure (TFS) and the slower changes in amplitude envelope (E). Mathematically, this way of partitioning known as Hilbert transform takes the form of $S(t) = \Sigma_{i=1}^{M} a_i(t) \cos (\theta_i(t))$, in which the original signal $S(t)$ is decomposed into $M$ sets of the cosine waves weighed by the instantaneous amplitude $a_i(t)$. Note that $\theta(t)$ is the instantaneous phase of the cosine function. Thus, in this context, E and TFS correspond to instantaneous amplitude $a(t)$ and instantaneous phase $\theta(t)$ of the cosine function, respectively. The perceptual significance of acoustic TFS and E has been studied in many laboratories (Smith et al., 2002; Xu and Pfingst, 2003; Zeng et al., 2004; Lorenzi et al., 2006; Fogerty, 2011; Wang et al., 2011, 2015; Apoux et al., 2013; Li et al., 2015; Qi et al., 2017). It is often assumed that E contributes to speech perception, whereas TFS contributes to music or tone perception as well as perception of interaural time difference (Smith et al., 2002; Xu and Pfingst, 2003, 2008).

The role of acoustic TFS in speech perception is still under debate. Several studies have tested the perception of TFS speech (Lorenzi et al., 2006; Lorenzi et al., 2009; Moore, 2008; Hopkins and Moore, 2009). The TFS speech is created with the formula: TFS-speech $= \Sigma_{i=1}^{M} \sqrt{a_i(t)^2} \cos (\theta_i(t))$, in which the E information is discarded by replacing it with the root-mean-square (RMS) of the E. Results from those studies have indicated that TFS conveys a fair amount of speech information, and that TFS information is particularly helpful for speech recognition in fluctuating noise (Hopkins and Moore, 2009). However, the acoustic TFS can be reconverted into neural E cues at the output of cochlear filters. Such E cues have been termed as reconstructed or recovered E information. Several recent investigations indicated that the reconstructed E may account for the intelligibility of the TFS speech (Lorenzi et al., 2012; Shamma and Lorenzi, 2013; Swaminathan et al., 2014; Léger et al., 2015). Thus, the interpretation of previous studies that have used TFS speech may have been confounded with the presence of recovered E information.

Most of the previous studies processed the speech material on the basis of the full length of the speech tokens or sentences. The present study adopted a method that

Author to whom correspondence should be addressed.
implemented short-time segment TFS processing. This approach was first proposed by Chen and Guan (2013) who used segment durations of 50 and 100 ms instead of the full length of the sentences. In the present study, we processed the TFS in six different short-time segments (i.e., 50 to 300 ms) and termed the resulting TFS as short-time temporal fine structure (STFS). By manipulating the segment duration, we could study the effects of rate of changes (20 to 3.33 Hz) in the amplitude on sentence recognition. Such a slow rate of changes in amplitude represents gross E information. Four experiments were carried out in the present study to examine the contributions of TFS as well as amplitude cues in short segments in sentence recognition and to evaluate potential benefit of using STFS in speech perception.

2. Method
2.1 Subjects
Fifty-two normal-hearing, native English-speaking young adults (21 males and 31 females) were recruited and randomly assigned to one of the four experiments with N of 14, 13, 13, and 12, respectively. The subjects were aged between 18 and 31 years old with a mean age of 24.6 years old. They were screened for normal hearing (≤20 dB hearing level) at octave frequencies between 250 and 8000 Hz. The use of human subjects was reviewed and approved by the Institutional Review Board of Ohio University.

2.2 Speech materials
The speech stimuli used were AzBio English sentences that consisted of 33 lists with 20 sentences in each list. The sentences ranged from 3 to 12 words (median = 7) in length. A list contained on average 140 words. All words were counted in scoring. All speech sentences were sampled at 22 050 Hz and were spoken by 2 male and 2 female adult speakers (Spahr et al., 2012).

2.3 Signal processing
Signal processing was performed in MATLAB (MathWorks, Natick, MA). To generate the short-time TFS (STFS), the entire sentence was first segmented into 50, 100, 150, 200, 250, or 300 ms. Each segment was then bandpass filtered into 8, 16, and 32 bands. With the overall frequency range of 64 to 8932 Hz, the bandpass filters were equivalent to 4-, 2-, and 1-ERB wide [ERB stands for the equivalent rectangular bandwidth and is an approximation to the bandwidths of the auditory filters of human cochlea (Glasberg and Moore, 1990)]. Hilbert transform was used to extract the instantaneous amplitude and phase of each band.

In experiment I, the amplitude information was removed by replacing it with 1. Thus, only the TFS of the original speech signal was maintained. The following equation shows the processing:

$$\text{FTFS}_j = \sum_{i=1}^{M} \cos (\theta_{ij}(n)), \quad j = 1, 2, \ldots, N,$$

where FTFS stands for flattened TFS, \(i\) represents the \(i\)th subband, and \(j\) represents the \(j\)th segment of the sentence processed. \(M\) is the number of bands that varied from 8 to 32 and \(N\) is the number of time segments each ranging between 50 and 300 ms.

In experiment II, the signal processing was similar to experiment I except that the instantaneous phase was replaced with random values so that TFS information was removed. The equation below shows the signal processing,

$$\text{STFS}_j = \sum_{i=1}^{M} \sqrt{a_{ij}(n)^2} \cos (\theta_{ij}(n)), \quad j = 1, 2, \ldots, N,$$

where \(i\) represents the \(i\)th subband and \(j\) represents the \(j\)th segment of sentence processed. As mentioned above, \(M\) was the number of bands that varied from 8 to 32 and \(N\) was the number of time segments of 50 to 300 ms each long. This was similar to the conventional way of processing TFS speech except that the processing was performed for each short-time segment, so that \(\text{STFS} = \{\text{STFS}_1, \text{STFS}_2, \ldots, \text{STFS}_N\}\).

In experiment III, the signal processing was similar to experiment II except that the instantaneous phase was replaced with random values so that TFS information was removed. The equation below shows the signal processing,
\[ RTFS_j = \sum_{i=1}^{M} \sqrt{a_{ij}(n)}^2 \cos(\text{rand}(\theta_{ij}(n))), \quad j = 1, 2, \ldots, N, \] (3)

where RTFS stands for randomized TFS and \( \text{rand}(\theta_{ij}(n)) \) is a random number generated to replace the phase value for the \( i \)th frequency band and at the \( j \)th time segment.

In experiment IV, the signal processing was similar to experiments II and III except that the instantaneous phase was replaced with sinusoids so that TFS information was removed. The equation below shows the signal processing,

\[ CTFS_j = \sum_{i=1}^{M} \sqrt{a_{ij}(n)}^2 \cos(2\pi f_i n + \varphi_{ij}), \quad j = 1, 2, \ldots, N, \] (4)

where CTFS stands for constant-frequency TFS and \( \varphi_{ij} \) is the random starting phase for the \( i \)th frequency band and at the \( j \)th time segment.

2.4 Procedure

Subjects listened under supra-aural headphones in a sound booth. The order of the 18 conditions (i.e., 3 bandwidths \( \times \) 6 segment lengths) was randomized. Each condition used a different AzBio sentence list. A total of 360 sentences were transcribed by each subject. The subjects listened to each sentence as many times as desired and then transcribed their answer. Prior to the real test, each subject went through a training period in which they listened to 72 processed sentences. Feedback was provided in practice but not in the real test.

3. Results and discussion

Figure 1 shows the average sentence-recognition performance as a function of the segment duration. The four panels of Fig. 1 represent data from the four experiments. A general linear model (GLM) analysis was performed to examine the effects of segment duration and processing bandwidth on sentence-recognition scores. The percent-correct scores of sentence recognition were treated as binomial data and a logit transformation of the percent-correct data was applied for the GLM analysis (Warton and Hui, 2011). Table 1 summarizes the results of the GLM analysis for the four experiments.

The four experiments yielded diverse results depending on the signal processing of the speech materials. In the following, the results are discussed in light of our current understanding of reconstructed E at the output of the cochlear filters from the TFS speech signals. Another aspect that we try to emphasize is the contribution of the amplitude information in the short-time segments to speech perception.

The TFS speech signals in experiment I were similar to those used in previous studies (Lorenzi et al., 2006; Swaminathan et al., 2014; Léger et al., 2015; Gilbert and Lorenzi, 2006). In those previous studies, Hilbert transform was used to decompose the speech signal (consonant tokens in a VCV format) into E and TFS components. The E information was removed and the TFS information of various numbers of bands were added together to form the TFS speech. The amplitude of the TFS was determined by the RMS values of the removed E. In experiment I of the present study, the amplitude of the TFS in all bands was set at 1. The short-time processing was irrelevant since the amplitude of all segments was set at the same level. The mean
sentence-recognition scores across all segment durations were approximately 9%, 45%, and 88% correct for bandwidths of 1, 2, and 4 ERBs, respectively. Our results were consistent with previous reports in that TFS speech is more intelligible if the bandwidth is wider from which the TFS speech is extracted using Hilbert transform. Recovered E information is found to be responsible for the results. The wider the bandwidth, the more readily the E will be recovered from the TFS signal (Ghitza, 2001; Zeng et al., 2004). Thus, experiment I of the present study provided additional evidence that recovered E contributes to sentence recognition in TFS speech.

Experiment II of the present study showed interesting interaction between bandwidth and segment duration. In this experiment, the amplitude of each segment was set at the RMS value of the E in that segment. This manipulation removed the detailed temporal fluctuation of the E but kept the overall amplitude contour of the sentence. The rate of amplitude changes is the reciprocal of the segment duration. At a bandwidth of 4 ERBs, the sentence-recognition scores were all above 90% correct. The recovered E was mostly responsible for such high performance. In the bandwidth conditions of 1 and 2 ERBs, the contributions of recovered E were reduced. On the other hand, the amplitude information for each short-time segment appeared to provide useful information for sentence recognition. The performance improved from approximately 13% and 53% correct at 300-ms segment for bandwidths of 1 and 2 ERBs, respectively, to nearly 100% correct at 50-ms segment (Fig. 1, second panel). Therefore, an amplitude cue at a rate of 20 Hz can provide perfect sentence recognition for TFS speech extracted from bands of 1- or 2-ERB wide. At a 10-Hz rate, sentence recognition was around 90% correct. Xu et al. (2005) reported that English vowel and consonant recognition saturated when E information was at or above 4 and 16 Hz, respectively. The present study indicates that good sentence recognition requires E information at or above 10 Hz. This result was also reminiscent of that of Saberi and Perrott (1999) who demonstrated that the recognition of locally time-reversed sentences deteriorated when the time segment was greater than 100 ms.

In experiments III and IV, the TFS information was removed and replaced with noise and sinusoids of constant frequencies, respectively. Both experiments produced very similar results. At 50-ms segment duration, amplitude information at a 20-Hz rate produced nearly perfect sentence recognition when the number of bands was 32 but only approximately 50% correct when the number of bands was 8. Sentence recognition performance dropped precipitously when the segment duration was 100 ms and further decreased to floor when the segment duration was 150 ms or longer. Note that the sentence-recognition performance at 8 bands of 50-ms segment duration was equivalent to that at 32 bands of 100-ms segment duration. This result provides evidence to a tradeoff between temporal and spectral cues for sentence recognition. Such a tradeoff has been previously demonstrated in English phoneme recognition as well as lexical tone recognition (Xu and Pfingst, 2008; Xu et al., 2002; Xu et al., 2005).

It is interesting to compare the speech recognition performance across all four experiments. Without amplitude information (experiment I), sentence recognition performance was approximately 9%, 45%, and 88% correct for bandwidths of 1, 2, and 4 ERBs, respectively. Adding amplitude information improved the performance dramatically (experiment II). Yet, with only amplitude and no TFS information (experiments III and IV), sentence recognition was very poor. We could not attribute the much better performance in experiment II entirely to the recovered E information. It appears that the TFS information facilitates speech perception in addition to the amplitude information. Whether the TFS and the amplitude information work independently or constructively remains to be tested (cf. Seldran et al., 2011).

In the present study, we manipulated the amount of TFS and amplitude information in speech signals using short-time processing. Results showed interesting

<table>
<thead>
<tr>
<th>Experiment</th>
<th>N</th>
<th>Segment duration</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(\beta)</td>
<td>t</td>
</tr>
<tr>
<td>I</td>
<td>14</td>
<td>-0.947</td>
<td>-7.5</td>
</tr>
<tr>
<td>II</td>
<td>13</td>
<td>-0.716</td>
<td>-70.7</td>
</tr>
<tr>
<td>III</td>
<td>13</td>
<td>-1.371</td>
<td>-76.6</td>
</tr>
<tr>
<td>IV</td>
<td>12</td>
<td>-2.468</td>
<td>-66.7</td>
</tr>
</tbody>
</table>
interactions among the time segment, TFS, amplitude, and recovered E on sentence recognition. While earlier studies illustrated the contribution of TFS information in speech perception (Lorenzi et al., 2006; Lorenzi et al., 2009; Moore, 2008; Hopkins and Moore, 2009, 2010; Gilbert and Lorenzi, 2006), several new studies suggested that recovered E in the TFS speech is responsible for the speech intelligibility (Swaminathan et al., 2014; Léger et al., 2015). The results of the present study suggest that we cannot completely rule out the contributions of TFS to speech perception. The results indicate that amplitude information in short-time segments facilitates speech perception when TFS information is available. Therefore, short-time processing of TFS speech is probably beneficial when implementing speech-processing strategies in hearing instruments such as hearing aids and cochlear implants for the hearing impaired. Whereas the ability to utilize TFS information in the hearing-impaired is a matter of debate (Moore, 2008; Hopkins and Moore, 2009; Léger et al., 2015), further testing of the short-time processed TFS speech is necessary. Another area that might show the beneficial effect of short-time processing of the TFS speech is speech perception of tonal languages such as Mandarin Chinese or music perception (Smith et al., 2002; Xu and Pfingst, 2003). Only one study (Chen and Guan, 2013) has tested Mandarin-Chinese speech perception using short time segments of 50 and 100 ms in TFS speech processing. The vowel duration in conversational speech of Mandarin Chinese is approximately 100 ms but is close to 200 ms in text reading (Yang et al., 2017). Therefore, future studies using various segment durations in TFS speech processing for Mandarin Chinese are warranted.

Acknowledgments

The authors are grateful to Kyle Brown, Ali Colopy, Samantha Cross, Yitao Mao, Lauren Muscari, and Jing Yang for their technical and editorial assistance. The study was supported in part by the National Natural Science Foundation of China (Grant No. 81300820/61071187) (L.H.), the Special Research Fund for the Public Welfare Industry of Health (Grant No. 201202001) (L.H.), and the NIH/NIDCD Grant No. R15- DC014587 (L.X.).

References and links


