Effects of envelope filter cutoff frequency on the intelligibility of Mandarin noise-vocoded speech in babble noise: Implications for cochlear implants

Guangting Mai, James W. Minett, William S-Y. Wang

Language Engineering Laboratory, Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong

{gtmai, jminett, wsywang}@ee.cuhk.edu.hk

Abstract

In cochlear implants, limited spectral and temporal information is provided. Previous studies argued for different effects of temporal information on speech identification in adverse environments. Particularly, it is unclear how speech intelligibility is influenced by the low-pass cutoff frequency of temporal envelope extractors in noise. The current study explored this issue with Mandarin noise-vocoder simulation in babble noise. Mandarin spoken sentences were mixed with multi-talker babble at different SNRs (4-, 6- and 8-dB) and then noise-vocoded. The vocoders had three numbers of logarithmic-spaced frequency channels (6, 8 and 10) with low-pass filtering the Hilbert envelope at 50 or 500 Hz in each channel. We measured the sentence intelligibility with native normal-hearing subjects and found that intelligibility for the high frequency cutoff condition (500-Hz) is significantly higher at 8-dB SNR with 6 and 8 channels, but significantly lower at 4-dB SNR with 6 channels than the low frequency cutoff condition (50-Hz). This finding suggests that high cutoff frequency of the envelope extractor in noise-vocoders can either improve or impair speech intelligibility in babble noise, hinging on the number of channels and the SNR level. Potential implications of the current finding for cochlear implant designs were further discussed in the paper.

Index Terms: noise-vocoder, envelope cutoff frequency, intelligibility, babble noise

1. Introduction

The temporal structure of speech consists of a wide range of frequency components. Rosen [1] defined three ranges of components to categorize speech information: (1) low-rate envelope fluctuates at frequencies below 50 Hz, corresponding to the rates of articulatory movement; (2) high-rate envelope (periodicity) fluctuates at frequencies in the range of 50 to 500 Hz, related to the periods of the fundamental frequencies ($F_0$); (3) fine structure fluctuates at frequencies above 500 Hz, corresponding to harmonic and detailed spectral structures. In many recent cochlear implant (CI) processing studies, the noise/tone-vocoder, which conveys the envelope of speech with fine structures replaced by noise or sinusoidal carriers, has been widely used to simulate the effects of CI processing. Xu and Zheng [15] suggested that relatively low low-pass cutoff frequency up to 16-Hz in noise-vocoders is sufficient for consonant and vowel recognition in noise. Since in CIs, low electrical stimulation rates, which was found to yield a better modulation detection threshold than high stimulation rates [16], is sufficient to represent envelope information with low-pass cutoff frequency, Xu and Zheng [15] suggested that relatively low low-pass, rather than high cutoff frequency of the envelope extractor may be a better choice for CIs in noise.

In previous CI simulation experiments with normal-hearing listeners, providing envelope information below 50 Hz in only a small number of vocoder frequency channels has been found to be sufficient for excellent intelligibility in quiet environments [2][3][5]. However, it is still unclear how speech intelligibility is influenced by the amount of temporal information being provided in terms of low-pass cutoff frequency of the envelope extractor in noisy environments. On one hand, relatively high cutoff frequencies were found to be important in concurrent competing sound environments. Stone et al. [7] demonstrated that low-pass cutoff frequency at 180 Hz led to significantly higher performances than cutoff frequency at 45 Hz in noise- and tone-vocoders for various SNR levels when normal-hearing listeners were asked to recognize English spoken sentences in the background of a single competing talker with different dynamic $F_0$ from the target talker. Hong and Turner [12] further found that high cutoff is significantly more helpful than low cutoff frequency in auditory stream segregation in CIs. These results are consistent with findings that $F_0$-related information provides critical cues for separating simultaneous sounds in human auditory scene analysis [13][14]. On the other hand, Xu and Zheng [15] showed that low-pass cutoff frequency up to 16-Hz in noise-vocoders is sufficient for consonant and vowel recognition in noise. Since in CIs, low electrical stimulation rates, which was found to yield a better modulation detection threshold than high stimulation rates [16], is sufficient to represent envelope information with low-pass cutoff frequency, Xu and Zheng [15] showed that even high low-pass, rather than high cutoff frequency of the envelope extractor may be a better choice for CIs in noise.
modulations than using low cutoff in noisy environments, which could adversely affect the speech intelligibility.

In the current study, we used three different numbers of channels in noise-vocoders and three SNR levels during the simulations with normal-hearing subjects (see details in Part 2).

2. Materials and signal processing

2.1. Signal processing procedure

Twenty native Mandarin normal-hearing subjects (graduate and undergraduate students from The Chinese University of Hong Kong, aged 19 to 30 yrs with 21 yrs in average) were recruited. They were required to listen to and identify the words and syllables contained in Mandarin sentences under the same babble noise background. Each sentence was a Semantically Unpredictable Sentence (SUS), naturally produced by a native male speaker (F₀s dynamically ranging from 80 Hz to 240 Hz averaged at 150 Hz) at around 4-Hz syllabic rate with each syllable having similar loudness. SUSs are sentences that are syntactically acceptable but semantically anomalous [19]. The purpose of using SUSs was to reduce the possibility that subjects guessed the content of each sentence on the basis of the semantic context, and forced the subjects to focus on the words they heard. Each SUS consisted of four disyllabic (2-character) words. All the words were manually chosen from the 9,000 most frequent items of a contemporary Mandarin word frequency corpus [20]. For instance, an SUS “地图突破严肃地点的” has four disyllabic (2-character) words “地图” (“map”), “突破” (“break through”), “严肃” (“serious”) and “地点” (“place”), while “的” is the particle without substantive meaning. All SUSs had similar duration of around 2,300 ms. The babble noise was created by mixing distinct Mandarin utterances at the same energy level spoken respectively by 5 male and 5 female native speakers, with all utterances extracted from the CUCall corpora (a collection of Mandarin telephone spoken language corpora) [21]. The 10 speakers all had different average F₀s ranging from 90 Hz to 280 Hz (dynamic ranges from 70 Hz to 410 Hz). In post-test runs after the main experiments, the resultant babble was confirmed to be unintelligible by all the subjects.

Each target SUS was first confined to the frequency range between 100 and 5500 Hz and mixed with the babble noise within the same frequency range at four SNRs of 4, 6, 8 and 30 dB before the noise-vocoding. In each mixture, the babble noise as the background started and ended respectively at around 1000 ms before the onset and after the offset of each sentence. The mixtures were then band-passed into multiple channels (6, 8, and 10), equally spaced in logarithmic scale, from 100 Hz to 5500 Hz. The numbers of channels were selected to reflect the typical numbers effectively used in cochlear implants [22]. The envelope of each channel was then extracted using the Hilbert Transform (HT) (Hilbert envelope) and low-passed with a zero-phase Hamming-based FIR filter. The filter response was -6dB at either 50 Hz (low cutoff, E filter) or 500 Hz (high cutoff, P filter). The envelope was subsequently used to modulate the noise carrier which was the Hilbert fine structure of the white Gaussian noise (also obtained through HT) in the same channel. The modulated signals in each channel were then band-passed again into the

2.2. Ranges of frequency channels

Table 1 shows the logarithmically-spaced distributions of the three channel conditions.

<table>
<thead>
<tr>
<th>condition</th>
<th>Chann. No.</th>
<th>Range (Hz)</th>
<th>Bandwidth (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-channel condition</td>
<td>1</td>
<td>100 - 149</td>
<td>49</td>
</tr>
<tr>
<td>2</td>
<td>149 - 223</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>223 - 333</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>333 - 497</td>
<td>164</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>497 - 742</td>
<td>245</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>742 - 1107</td>
<td>365</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1107 - 1653</td>
<td>546</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1653 - 2468</td>
<td>815</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2468 - 3684</td>
<td>1216</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3684 - 5500</td>
<td>1816</td>
<td></td>
</tr>
</tbody>
</table>

3. Tasks and experiment procedure

We created a set of 120 SUSs in different contents for the formal testing session, in which there were 24 conditions (3 channels × 4 SNRs × 2 filters) and 5 different SUSs in each condition. All subjects were presented with this same set of SUSs. To reduce the possible biases caused by the contents of the sentences (e.g., different word frequencies), subjects were divided into 5 groups (4 subjects each) and in each condition different groups were presented with different subsets of SUSs. This was achieved through the signal processing (Part 2.1) preparing the sentence stimuli prior to the experiments. During the audio presentation, the SUSs were divided into 5 successive blocks (24 SUSs each) and SUSs with the same condition were presented separately in different blocks.

Subjects were seated in a quiet speech-recording room in the DSP and Speech Technology Laboratory of the Dept. of Electronic Engineering at The Chinese University of Hong Kong. The stimuli were presented diotically via a high-quality
Sennheiser HD-419 headphone with a fixed sound level reported comfortable by every subject. Subjects were instructed to write down the words or syllables they heard in each SUS on an answer sheet after it was played. To minimize the homophone confusions, they were required to transcribe what they heard using “Pinyin” (the official phonetic alphabet system transcribing pronunciations of Chinese characters into Latin scripts) if they could only recognize one syllable within a word. Before the formal test, each subject attended a 20- to 30-minute training session consisting of an extra set of 20 SUSs (all different from those in the formal test). Each training SUS was presented three times with feedback of the correct answer so that subjects could get familiar with the experiment beforehand. In the formal test, each SUS was played only once with no feedback provided. The entire experiment lasted for around 1.5 hours.

4. Results

The speech recognition performance was calculated in syllable identification accuracies, which refers to the percentage of correct syllables recognized by the subjects. Syllable identification accuracies across the 20 subjects were illustrated in Fig. 1, as a function of SNR (4, 6, 8 and 30 dB) and Filter (E/P) for the 6-, 8- and 10-channel conditions, respectively. Black and white bars represent the conditions of the E and P filter, respectively.

A three-way Within-Subject (Repeated Measures) ANOVA shows that there is no significant interaction between Channel, SNR and Filter (F(6,19)=0.860, p=0.503). There are significant interactions between Channel and SNR (F(6,19)=6.509, p<0.0001) and between SNR and Filter (F(3,19)=7.626, p=0.003) but no interaction between Channel and Filter (F(2,19)=1.211, p=0.306). There are significant main effects of Channel (F(2,19)=172.94, p<10^-16) and SNR (F(3,19)=281.321, p<10^-30) but marginally significant effect of Filter (F(1,19)=4.11, p=0.057).

To concretely assess the effects of envelope low-pass cutoff frequency in various SNR levels with different frequency channels, post hoc pairwise analysis were implemented comparing the differences between the E and P filters. P values of each pairwise comparison are shown right above the black and white bars in Figure 1. In the 6-channel condition, performances with the E and P filters were distinct across different SNRs: in the case of 30-dB SNR, which was set to determine whether high-rate envelope (50-500 Hz) are still important in a near-clean environment, no significant difference was found between the two filters (p>0.4); the P filter led to a significantly better performance than the E filter at 8-dB SNR (p<0.001) whilst it was the other way round at 4-dB SNR (p=0.002). In the 8-channel condition, the performance with the P filter was significantly better than that with the E filter at 8-dB SNR (p=0.03), but no significant differences between the two filters were found at 4-, 6- and 30-dB SNRs. In the 10-channel condition, no significant differences between the two filters were found at any of the SNR levels.

In summary, the results show that the effect of Filter for syllable accuracies is modulated by both the number of channels and the SNR level. Pairwise comparisons show that significant effects of Filter were only found in the 6- and 8-channel conditions but not in the 10-channel condition. High low-pass cutoff frequency (500 Hz) was significantly more beneficial at relatively high SNR (8-dB) in the 6- and 8-channel conditions and on the other hand, however, significantly more detrimental at relatively low SNR (4-dB) in the 6-channel condition for the speech intelligibility than the low low-pass cutoff frequency (50 Hz).

5. Discussions

We investigated the effects of low-pass cutoff frequency of temporal envelope extractors in each frequency channel for the intelligibility of Mandarin noise-vocoded spoken sentences in babble noise with native normal-hearing listeners. The results show that the effects depends on both the numbers of channels and the SNR level:

1) In the near-clean environment (30-dB SNR), high low-pass cutoff frequency (500 Hz) did not significantly improve intelligibility compared to low cutoff frequency (50 Hz). This is inconsistent with previous studies which showed that high-rate envelope information are important for the intelligibility of Mandarin noise-vocoded speech due to its importance for the lexical-tone identification in Mandarin [4][6][17]. This could be because, in previous studies, either low spectral resolutions [17] or isolated syllable stimuli [4][6] were used, whereas the current study used entire sentences as stimuli. The spectral resolutions with 6 to 10 channels here conserved high degrees of segmental (consonants and vowels)
information for identification, which partially compensated for the loss of high-rate envelope information for tone identification (in current SUs, all the words are real and valid disyllabic words, correct identification of segments within a word will help in tone identification of the syllables within the same word).

(2) The high low-pass cutoff frequency (500 Hz) was found to be significantly higher for intelligibility at 8-dB SNR with 6 and 8 channels but became significantly lower than the low cutoff frequency (50 Hz) when the SNR was decreased to 4-dB with 6 channels. No effects of low-pass cutoff frequency were found at any SNR with 10 channels.

These results prove our speculations that that using high envelope cutoff frequency may introduce more noise modulations than using low cutoff in noisy environments, which adversely affects the speech intelligibility, especially at the low SNR level (4-dB), although high-rate envelope information has been proved to be important for tonal languages like Mandarin (also see the discussions above). The advantageous effects of higher cutoff frequency occurred only at the relatively high SNR level (8-dB).

The results also show that the effects of envelope low-pass cutoff frequency occurred with 6 and 8 channels, but did not occur with 10 channels. This can be reflected by the bandwidth of each frequency channel when the vocoder is created with relatively high numbers of channels (10-channel, see Table 1). Seen from Table 1, for 10-channel condition, bandwidths of most of the channels are narrower than 1000 Hz (twice of the 500 Hz). To allow the modulations at the high low-pass cutoff frequency (500 Hz) to be passed at full depth, the channel bandwidth should be at least twice of such cutoff frequency. Therefore, although the high low-pass cutoff frequency was used, the high-rate modulations (50 – 500 Hz) were not totally preserved since the spectral sidebands produced by such envelope modulations were attenuated by the narrow bandwidths of the 10-channel noise vocoders. For the 6- and 8- channel conditions, however, such attenuation effects were smaller.

In summary, our study demonstrates that relatively high low-pass cutoff frequency of envelope extractors in Mandarin noise-vocoded speech in multi-talker babble can either improve or impair the intelligibility, depending on both the number of channels and the SNR level. Since we used noise vocoder simulations with normal-hearing subjects to assess the effects of available acoustic information in the processed signals, our results are indicative for most CIs today that mainly envelope information are taken advantage of. High-rate envelope information have recently been suggested to be included in CI processing in complex environments [12]. Therefore, our current results argue for a different scenario against this necessity and are suggestive to the designs of CIs that using high low-pass cutoff frequency of envelope extractors is not consistently optimal in complex environments.

The main implication of this study is that in order to maximize the sentence intelligibility in multi-talker babble environments in CIs with control of the low-pass frequency of envelope extractors, high cutoff frequency should be used at high SNR levels. On the other hand, low cutoff frequency should be used at low SNR levels, especially when the number of frequency channels is relatively low. The turning point of the SNR levels should be carefully examined in real applications. However, it should be also suggested that the current speculation requires testing with CI subjects.

6. References

