Spectral density affects the intelligibility of tone-vocoded speech: implications for cochlear implant simulations

Stuart Rosen, Yue Zhang and Kathryn Speers

Speech, Hearing and Phonetic Sciences, University College London

Chandler House, 2 Wakefield Street, London WC1N 1PF, United Kingdom

stuart@phon.ucl.ac.uk, yue.zhang.12@ucl.ac.uk, k.speers.12@ucl.ac.uk

Sparse spectra harm speech intelligibility

Abstract

For small numbers of channels, tone vocoders using low envelope cutoff frequencies are less intelligible than noise vocoders, even though the noise carriers introduce random fluctuations into the crucial envelope information. Here it is shown that using tone carriers with a denser spectrum improves performance considerably over typical tone vocoders, at least equalling, and often surpassing, the performance possible with noise vocoders. In short, the spectral sparseness of tone vocoded sounds for low channel numbers, separate from the degradations introduced by using only a small number of channels, is an important limitation on the intelligibility of tone-vocoded speech.

PACS numbers: 43.71.Es, 43.71.Ky, 43.66.Ts

1. Introduction

Noise- and tone-vocoded speech have been widely used to investigate the role of temporal cues in speech perception and simulate the information in speech received by users of cochlear implants (CIs). Typically, the speech signal is first divided into a certain number of frequency bands and the amplitude envelope of each band is extracted to modulate a carrier signal that is either a tone at the centre frequency of the band or wide-band noise which is then filtered to the analysis bandwidth. Various parameters of the vocoder significantly affect the intelligibility of the resulting signal. For instance, the number of bands constrains the spectral resolution and dynamics of the speech, and the degree of envelope smoothing (usually implemented through lowpass filtering) constrains the range of temporal information transmitted [Loizou et al., 1999] [Roberts et al., 2011] [Rosen, 1992] [Xu and Pfingst, 2008].

The carrier signal, in comparison, has a less obvious impact. Dorman et al. (1997) reported no differences in performance between noise and tone carriers for vowel, consonant and sentence recognition. Whitmal et al. (2007), on the other hand, generally found better speech recognition for a tone carrier, at least when using noise carriers of the type typically used in such studies. Souza and Rosen (2009) identified an interaction between carrier type and cutoff frequency: stimuli constructed with tone carriers were less intelligible than noise-vocoded sounds for a low envelope cutoff frequency (30 Hz), but more intelligible for a high envelope cutoff (300 Hz). Note too that Whitmal et al. (2007) used a high envelope cutoff frequency (300 Hz but limited to half the analysis bandwidth), as did Dorman et al. (1997) (400 Hz).

The ambiguous effect of carrier type may be due to multiple reasons. Higher envelope cutoff frequencies on tone-vocoded speech provide both better temporal information and access to intonation through signalling the F0 contour [Rosen, 1992]. This benefits tone-vocoded speech in more than one way. Also, there is no inherent fluctuation in a tone carrier as compared to a noise carrier, which further benefits tone-vocoded speech when compared to its counterpart [Souza and Rosen, 2009] [Whitmal et al., 2007].

Another possible explanation lies in the differences in spectral density for stimuli used in these studies. For envelope cutoff frequencies sufficiently lower than the fundamental frequency (F0), each band of tone-vocoded speech is effectively composed of only one amplitude modulated tone. When the cutoff frequency is higher than F0, within each band there emerge sidebands that are the sum and difference frequencies between the carrier tone and F0 as well as all components in the low-pass filtered envelope. As a result, the spectrum is denser, especially when the number of bands in the vocoding is low. The spectral density of noise-vocoded speech, in comparison, is unaffected by the envelope cutoff frequency, since energy exists at every frequency in each band (Fig. 1). It is possible that the denser spectrum will contribute to higher speech intelligibility in two ways. First, the temporal information in each channel may be improved by having a broader range of frequencies modulated by the same envelope. Secondly, the spectral shape of the processed speech with denser spectrum will be a closer match to the spectral shape of the original speech. Therefore, the differences in results across study may arise from the use of different lowpass envelope cutoffs that leads to different spectral densities. The 30 Hz cutoff used by Souza and Rosen (2009) left the spectrum of tone-vocoded speech with 'holes' between each band that damaged its intelligibility. The 300 Hz cutoff, on the other hand, provided denser spectra than the 30 Hz cutoff tone-vocoded speech but sparser spectra than noise-vocoded speech. However, the inherent fluctuation of the noise carrier may interfere with the full use of temporal information [Gonzalez and Oliver, 2005], making it less intelligible than the tone carrier with 300 Hz cutoff. This was consistent with Whitmal et al. (2007), where the cutoff frequencies around 300 Hz similarly benefited the tone carrier more than noise carrier. Dorman et al. (1997)'s work applied different envelope cutoff for tone-vocoded (400 Hz) and noise-vocoded speech (160 Hz), making the result hard to interpret. Nevertheless, the high envelope cutoff for the tone vocoder would be expected to lead to higher levels of performance than low envelope cutoffs, so the denser spectrum may have led to performance equivalent to that obtained from noise vocoders. However, no previous studies can directly demonstrate the role of spectral density on speech intelligibility, due to the various confounds. Therefore, it is not clear what role spectral density actually plays and how much it contributes to the intelligibility difference for different types of vocoded speech.

Hence here, in two studies, we investigated the impact of spectral density on intelligibility, by comparing typical tone-vocoded speech to a denser tone-vocoded speech that has more spectral components in its spectrum without introducing more temporal and spectral information. As the effects of spectral density are only likely to be found for low band numbers, materials were chosen that allowed reasonable performance with small numbers of bands. One set of tests used everyday material of simple sentences in an open-set response format vocoded to 3 ,4 and 5 bands. Another test, one using a closed-set response, allowed for even fewer bands to be used (2, 3 and 4).

2. Experiment

2.1. *Participants*

12 and 10 self-reported normal-hearing native standard Southern British English speaking adults were recruited for the closed-set and open-set sentence tests respectively. All were students from UCL, ranging in age from 21 to 35 years (mean = 26 years). None of them had prior experience with vocoded speech. All participants consented to take part by reading and signing a consent form, as approved by the UCL Research Ethics Committee.

2.2. *Stimuli*

The stimuli for the closed-set response test were adapted from the Coordinate Response Measure sentence corpus (CRM) [Bolia et al., 2000]. Sentences took the form of 'Show the (animal) where the (colour) (number) is' (e.g. 'show the pig where the pink five is'). There were 6 possible animals (cat, cow, dog, duck, pig and sheep), 6 possible colours (black, blue, green, pink, red and white) and 8 possible numbers (1 to 9 excluding the bisyllabic 7). Sentences were recorded at a sampling frequency of 22.05 kHz from 4 different native standard Southern British English talkers (2 male and 2 female) in an anechoic chamber. Stimuli for the open-set sentence test consisted of BKB sentences [Bench et al., 1979], spoken by 2 standard Southern British English talkers (1 male and 1 female), also sampled at 22.05 kHz. Each of these sentences contained 3 key words, upon which the scoring was based.

For the typical tone- and noise- vocoded conditions, each file was digitally filtered into 2 to 6 bands, using sixth-order Butterworth IIR filters. Filter spacing was based on equal basilar membrane distance [Greenwood, 1990] across a frequency range of 70-6000 Hz. The output of each band was full-wave rectified and low-pass filtered (fourth-order Butterworth) at 30 Hz, using a zero-phase forwards-backwards technique to extract the amplitude envelope. The envelope was then multiplied by a carrier, either a tone at the band centre frequency or a noise. The resulting signal was filtered using the same bandpass filter as the first filtering stage. The rms level of the output signal was adjusted to match the original analysis band and the signal was summed across all bands.

For the dense tone-vocoded speech, the processing was the same as for typical tone-vocoded speech, except for the carrier. Here, the carrier was a set of 24 equal-amplitude sinusoidal components based on the centre frequencies of filters that would be equally spaced on the basilar membrane across the frequency range of 70 - 6000 Hz. Unlike typical tone-vocoded speech, in which each tone would be modulated by a different envelope, here all components within each filter band were modulated by the same envelope. This was created by synthesising for each filter band a distinct carrier wave that contained only those components of the full 24 that were within the analysis band cutoffs. For example, for a two-band vocoder, there was 12 components modulated by the envelope in the lower frequency band (86 - 960 Hz) and 12 modulated by the envelope in the higher frequency band (1124 - 5594 Hz) (Fig. 2). All stimuli were processed using MATLAB software.

Each test consisted of 9 conditions: three carrier types (noise, tone and dense tones) at each of three different numbers of bands (2, 3, 4 for the closed-set test and 3, 4, 5 for the open-set). Training stimuli were composed of a higher number of bands (5 and 6) to facilitate adaptation.

2.3. *Procedure*

Stimuli were presented to the participants binaurally through Sennheiser HD25 headphones at 66 dB SPL in a quiet room. All testing was conducted in MATLAB using special-purpose software controlled from a PC.

In the closed-set test, each participant was firstly trained with 72 sentences, randomly selected from 12 different sentence lists vocoded by different combinations of carrier signals and band numbers (2 to 6). The first 36 sentences only used 4-6 bands, whereas the second set used 2-4 bands. They were then tested with another 72 different sentences, 8 for each of the 9 conditions. Sentences were chosen randomly from each of the 4 talkers, but constrained by the condition that each talker was heard an equal number of times in each test or practice session. During both training and testing, participants were required to click the target number and colour they had heard on a response screen and were presented with visual feedback (smiley face for a correct response and sad face for an incorrect response). All responses were saved and the program proceeded to the next trial.

In the open-set sentence test, participants were trained with 36 sentences randomly selected from the test set (3, 4 and 5 bands). In each trial, participants were scored by the experimenter based on the number of keywords correctly reported. If the score was lower than 3, the unprocessed and processed stimuli were played twice as auditory feedback. Testing used another 180 sentences, 20 for each of the 9 conditions, and followed the same procedure only without feedback. Again, sentences were chosen randomly from each of the 2 talkers, constrained by the condition that each talker was heard an equal number of times in each test or practice session.

2.4. *Results*

Results of the closed-set response test are shown in Fig. 3. Participants' responses were fitted with mixed-effects logistic regression models, using the lme4 package [Bates et al., 2014] in R. Factors were entered into the model in the sequence below: the model firstly started with taking *listener* and *talker* as random intercepts; fixed effect factors *carrier*, *band number* (transformed to a logarithmic scale [Shannon et al., 2004]) and their interaction were then entered; finally random slopes were entered into the model. Factors were retained in the model only if they significantly improved the model fitting, using Chi-squared tests based on changes in deviance. The best model obtained was: *response* = (1 |*listener*) + *band number* + *carrier* (the term (1 |*listener*) indicates only a significant random intercept of *listener*, suggesting that there is a significant difference among listeners' average responses but not in the relationship between their responses and fixed effect factors). Both main factors were found significant (*band number*:χ^2=141.57, df=1, p<0.01; *carrier*:χ^2=35.57, df=2, p<0.01), with no significant interaction (χ^2=1.95, df=2, p>0.05). *Post hoc* analysis using Wald tests showed that correct responses increased with the number of bands (z=11.34, p<0.01). For carriers, dense tones were significantly better than tones (z=-5.73, p<0.01) but not noise (z=1.90, p=0.06), and the noise carrier was better than the tone carrier (z=-3.94, p<0.01).

The results for the open-set sentence test are given in Fig. 4. The proportion of correct responses was fitted using the same procedure and the best model was also: *response* = (1 |*listener*) + *band number* + *carrier*. *Band number* (χ^2=415.00, df=1, p<0.01) and *carrier* (χ^2=493.31, df=2, p<0.01) were significant but there was no significant interaction (χ^2=3.62, df=2, p>0.05). More bands resulted in higher intelligibility (z=20.52, p<0.01). The dense tones led to significantly better performance than tones (z=-21.03, p<0.01) and noise (z=-0.805, p<0.01), and the noise carrier was better than the tone carrier (z=-13.81, p<0.01).

3. Discussion

This study clearly demonstrates that increasing spectral density alone can enhance the intelligibility of tone-vocoded speech significantly without providing any extra spectral or temporal information. Dense tone-vocoded speech in this experiment has the same degree of spectral resolution and the same modulations as in typical tone-vocoded speech, only with more sinusoidal components. For two types of speech material and across all number of bands, dense tone-vocoded speech was significantly more intelligible than normal tone-vocoded speech. This confirms what has only been suggested in past studies concerning the positive impact of a denser spectrum on intelligibility [Assmann and Nearey, 2008] [Churchill et al., 2014] [Souza and Rosen, 2009].

Other findings were also consistent with previous studies. The noise carrier was found to be more intelligible than a tone carrier that has a much sparser spectrum. However, when the spectra were similarly dense (dense tone- and noise-vocoded speech), the inherent fluctuations of the noise presumably disrupted the temporal envelope information to some degree, thereby giving poorer performance in the open-set sentence test. Dense tone-vocoded speech was not significantly better than noise-vocoded speech in the closed-set response test, possibly due to the small amount of testing.

The disadvantage of a sparse spectrum, as in typical low envelope cutoff tone-vocoded speech, may lie in the incomplete use of the available spectral and temporal cues. With spectral 'holes' between each sinusoidal carriers, it may be hard for listeners to fuse disparate frequency components into a single auditory object and observe the between-band amplitude changes that convey information about spectral dynamics [Roberts et al., 2011]. Furthermore, the temporal envelope of each band may be underrepresented with one sinusoid. Therefore, having the same temporal information available in more frequency components, ideally without random fluctuation, may optimise the representation of aspects of speech essential to intelligibility [Roberts et al., 2011] [Rosen and Iverson, 2007].

This study has important implications for experiments using vocoded speech as CI simulation. The choice of carrier signal affects both the type and amount of acoustic information transmitted and these effects interact with the choice of envelope cutoff frequency. Considering that vocoded speech has been used in a wide range of research topics, including speech perception plasticity, effectiveness of CI parameter settings and training schemes [Guediche et al., 2014] [Shafiro et al., 2012], it is essential to take into account the impact of the various parameters in vocoder processing, in order to obtain a clear picture of the experimental factors of main interest.

Acknowledgements

Some of the work reported here was performed as part of the UCL Speech and Language Sciences MSc thesis of Kathryn Speers. Yue Zhang is sponsored by a student scholarship from the CSC (Chinese Scholarship Council).

References and links

Assmann, P. F., & Nearey, T. M. (**2008**). "Identification of frequency-shifted vowels," The Journal of the Acoustical Society of America. 124(5), 3203-3212.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (**2014**). "lme4: Linear mixed-effects models using Eigen and S4," R package version 1.1-7, http://CRAN.R-project.org/package=lme4. (Last viewed 19 Aug. 2014.)

Bench, J., Kowal., A., & Bamford, J. M. (**1979**). “The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children,” British Journal of Audiology. 13(3), 108-112.

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (**2000**). “A speech corpus for multitalker communications research,” The Journal of the Acoustical Society of America. 107(2), 1065-1066.

Churchill, T. H., Kan, A., Goupell, M. J., Ihlefeld, A., & Litovsky, R. Y. (**2014**). “Speech Perception in Noise with a Harmonic Complex Excited Vocoder,” Journal of the Association for Research in Otolaryngology. 15(2), 265-278.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). “Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs,” The Journal of the Acoustical Society of America. 102(4), 2403-2411.

Gonzalez, J., & Oliver, J. C. (**2005**).“Gender and speaker identification as a function of the number of channels in spectrally reduced speech,” The Journal of the Acoustical Society of America. 118(1), 461-470.

Greenwood, D. D. (**1990**).“A cochlear frequency position function for several species29 years later,” The Journal of the Acoustical Society of America. 87(6), 2592-2605.

Guediche, S., Holt, L. L., Laurent, P., Lim, S. J., & Fiez, J. A. (**2014**). "Evidence for cerebellar contributions to adaptive plasticity in speech perception," Cerebral Cortex. bht428.

Loizou, P. C., Dorman, M., & Tu, Z. (**1999**). "On the number of channels needed to understand speech," The Journal of the Acoustical Society of America. 106(4), 2097-2103.

Roberts, B., Summers, R. J., & Bailey, P. J. (**2011**). "The intelligibility of noise-vocoded speech: spectral information available from across-channel comparison of amplitude envelopes," Proceedings of the Royal Society B: Biological Sciences. 278(1711), 1595-1600.

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.

Rosen, S., & Iverson, P. (**2007**). "Constructing adequate non-speech analogues: what is special about speech anyway?" Developmental Science. 10(2), 165-168.

Shafiro, V., Sheft, S., Gygi, B., & Ho, K. T. N. (**2012**). "The influence of environmental sound training on the perception of spectrally degraded speech and environmental sounds," Trends in amplification. 16(2), 83-101.

Shannon, R. V., Fu, Q. J., & Galvin 3rd, J. (**2004**). "The number of spectral channels required for speech recognition depends on the difficulty of the listening situation," Acta Oto-laryngologica. Supplementum. 552, 50-54.

Souza, P., & Rosen, S. (**2009**). "Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech," The Journal of the Acoustical Society of America. 126(2), 792-805.

Whitmal III, N. A., Poissant, S. F., Freyman, R. L., & Helfer, K. S. (**2007**). "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," The Journal of the Acoustical Society of America. 122(4), 2376-2388.

Xu, L., & Pfingst, B. E. (**2008**). "Spectral and temporal cues for speech recognition: implications for auditory prostheses," Hearing Research. 242(1), 132-140.

Figure captions

Figure 1: Spectrograms of the sentence 'The birch canoe slid on the smooth planks' uttered by a male speaker, vocoded with different combinations of carrier signals and envelope cutoff frequencies. The left panel shows the four-band noise-vocoded versions, with the top one having an envelope cutoff frequency of 30 Hz (4n30) and the bottom one a cutoff frequency of 300 Hz (4n300). Correspondingly, the right panel demonstrates the four-band tone-vocoded versions, with the top one a cutoff frequency of 30 Hz (4s30) and the bottom one a cutoff frequency of 300 Hz (4s300).

Figure 2: Spectrograms of the two-band tone-vocoded sentence 'Show the cat where the blue six is' uttered by a female speaker, with different types of carrier signals. (a) shows the dense tone carrier used in the experiment (Mm. 1). The carrier is a set of 24 sinusoids that are of the same amplitude, equally distant on the basilar membrane. The lower 12 components (band 1) are modulated by the same amplitude envelope and the upper 12 components (band 2) are modulated by another. In comparison, (b) shows the tone carrier, which has only two sinusoidal components(Mm. 2). Although the tone in band 1 is modulated by the same envelope as in band 1 of (a) and the tone in band 2 by the same envelope as in band 2 of (a), the spectral density of the two sounds are different.

Mm. 1. Sound file of the two-band dense tone-vocoded sentence. This is a file of type

‘.wav’ (89 Kb).

Mm. 2. Sound file of the two-band typical tone-vocoded sentence. This is a file of type ‘.wav’ (89 Kb).

Figure 3: Boxplots of proportion correct for the closed-set response test as a function of band numbers, with carrier type indicated by differently shaded boxes.

Figure 4: Boxplots of proportion correct for BKB sentences in the open-set test as a function of band numbers, with carrier type indicated by differently shaded boxes.