Acta Acustica united with Acustica
Vol. 104 (2018) 835 – 838

DOI 10.3813/AAA.919253

The Factor Analysis of Speech: Limitations and Opportunities for Cochlear Implants

Jacques A. Grange, John F. Culling
School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, U.K.
grangeja@cf.ac.uk

Summary
Current spread is known to limit the number of independent spectral channels in cochlear implants. The outcome of an experiment employing cochlear implant simulations indicated that current spread is not the only limitation on the benefit of increasing the number of electrodes: for both sentences and digit triplets, improvements in speech reception threshold slowed markedly once more than seven electrodes/processing channels were simulated. Factor analysis of speech envelopes from the output of an auditory filterbank confirmed that speech contains 6-8 independent sources of information, causing finer spectral filtering to produce redundant information in adjacent channels. It is possible that factor analysis can be used to refine the frequency maps used in cochlear implants in order to minimise the effects of current spread.

PACS no. 43.66.Dc, 43.66.Ts, 43.71.An

1. Introduction

Current spread limits the benefit of increased numbers of electrodes in cochlear implants, because ganglion cells at a given location on the spiral ganglion are stimulated by multiple electrodes (e.g. Abbas *et al.* [1]). Friesen *et al.* [2] showed that the percent-correct sentence recognition in noise of cochlear implant users plateaued once about seven electrodes were activated. They attributed this effect to the influence of current spread. In order to research this effect, Grange *et al.* [3] developed a novel acoustic simulation of a cochlear implant, the SPIRAL vocoder. SPIRAL modulates a large, and fixed, number of sinusoidal carriers according to the mixed influences of an independently controlled number of electrodes. Using SPIRAL, Grange *et al.* confirmed a limiting effect of current spread on the benefit of additional electrodes. However, in the present study, we use SPIRAL to show that the increasing benefit of additional electrodes shows a marked inflection at around seven electrodes even in the absence of current spread, indicating the existence of a more fundamental limiting factor.

We postulated that this second limiting factor might be informational redundancy in speech itself. In order to investigate this possibility, we factor analysed speech envelopes extracted from an auditory filterbank. Factor analysis allows one to quantify the number of independent modulators in the speech signal and, through the factor loadings, to characterise the spectral extents of those modulators. These loading patterns indicated that conventional logarithmic spacing of analysis channels may not provide an optimal frequency map for cochlear implant processors.

2. Methods

2.1. Materials and vocoding

Speech and noise were mixed and vocoded using SPIRAL [3], a vocoder designed to simulate listening through a cochlear implant (CI) with normally-hearing listeners. In one experiment, the target speech consisted of IEEE sentences (Rothauser *et al.*, 1969) from the M.I.T. recordings, spoken by a male speaker (‘DA’). In the second experiment, digit triplets were used, spoken by a female and recorded in our laboratories. Each set of target material was mixed with speech-shaped noise, which was spectrally filtered to match the long-term spectrum of the target speech for that experiment.

SPIRAL employed 80 sinusoidal carriers equally distributed along an ERB scale in the 20–20000 Hz range. Input signal analysis was performed by rectangular bandpass 512-point finite impulse-response filters uniformly distributed along an ERB scale. The bandpass filters covered a 120–8658 Hz frequency range fully and without overlap, such that filter widths increased as the number of activated channels decreased. To extract temporal envelopes, the filtered waveforms were half-wave rectified and low-pass filtered with a 50 Hz cut-off. The centre fre-

Received 28 February 2018,
accepted 10 July 2018.

quency of a band was the place frequency of the corresponding simulated CI electrode. With no simulated current spread (spread of excitation set to -200 dB/oct), the temporal envelope extracted within a band modulated only a small number of tone carriers, whose frequencies were closest to the place of the simulated electrode.

2.2. Procedure

SRTs were measured using a one-up/one-down adaptive tracking method that kept the combined level of speech and noise at 65 dB A. For the IEEE sentences (that contain five key words), the adaptive track converged on the 30% point in the psychometric function by increasing the signal-to-noise ratio (SNR) by 2 dB if fewer than two key words in a given sentence were identified and otherwise decreasing it by 2 dB. For the digit triplets, the adaptive track converged on the 50% point by increasing the SNR by 2 dB if fewer than 2 digits were correctly reported, and decreasing it by 2 dB if not. The IEEE-sentence SRT measurements started with a low SNR (set 12 dB lower than SRTs measured in practice runs); the SNR was then increased in 4 dB steps until at least one word from the first target sentence was correctly identified, at which point the adaptive phase started. For digit triplets, the initial SNR was high (set 12 dB above practice-run SRTs) and was decreased in 4 dB steps until less than 2 digits were correctly identified. Sentence SRTs employed lists of 10 sentences and were computed as the mean of the last 8 computed SNRs. Digit-triplet SRT tracks stopped after 10 reversals and SRTs were computed as the mean of all the computed SNRs of the adaptive phase. Each SRT condition had a different number of activated channels. These were 4, 5, 6, 8, 10, 15 or 20 for IEEE-sentences and 3, 4, 6, 8, 10, 15 or 20 for digit-triplets. For sentence SRTs, the sentence order was fixed and the condition order was quasi-randomized and rotated against the material. 210 sentences were used in total. The experiments included conditions with current spread that are not reported here. Each participant performed four practice SRT runs prior to the experimental runs.

For each experiment, 21 young adults participated, recruited from the Cardiff University undergraduate population and self-reported as normally hearing (age mean 19, range 18–21). Briefing, consent and debriefing followed the rules set out by the institutional review board.

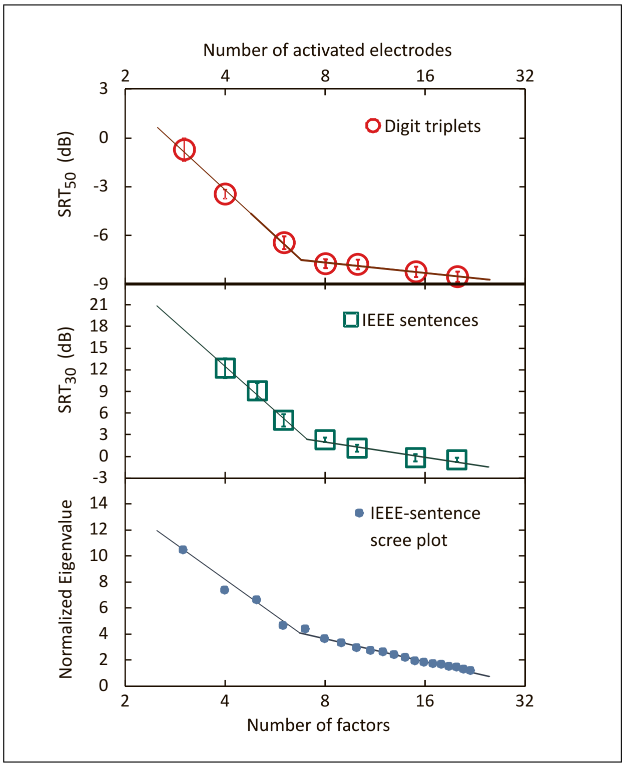

Figure 1. Top two panels: SRTs for digit triplets and IEEE sentences heard through the SPIRAL CI simulator without current spread simulation. Bottom panel: scree plot from the factor analysis of speech modulations in the IEEE sentences spoken by M.I.T. talker ‘DA’. The lines going through the data or screen-plot points are best bilinear fits. Error bars are standard error of the means.

2.3. Factor analysis

Factor analysis (FA) was conducted on the concatenation of all IEEE sentences spoken by the M.I.T. speaker DA, in the 200–8000 Hz range. The FA followed the procedure used by Ueda and Nakajima [4]: temporal modulation envelopes were extracted (by gammatone filtering operated every ¼ ERB, half wave rectification and low-pass filtering at 50 Hz), then squared and converted to z-scores such that a co-modulation analysis across frequency bands would be operated on a correlational basis. The resulting power envelopes were fed through a principal component analysis (PCA Matlab program from Brian Moore, University of Michigan, Nov. 2016), which operated a singular value decomposition of the power envelope matrix and allowed the user to specify the number of retained components. The derived component loadings were then subjected to a varimax rotation to maximize factor independence and obtain factor loading curves. Each factor loading curve represents a co-modulating region of the speech spectrum that modulates most independently of the others. In order to generate a “scree” plot of reducing normalized Eigenvalue as a function of the number of retained factors, normalized Eigenvalues were derived for each number of retained factors, through matrix multiplication of the normalized Eigenvectors and the z-scored power-envelope correlation matrix, followed by point-by-point division by the normalized Eigenvectors. The resulting scree plot represents the amount of information each additionally retained factor adds in the explanation of the temporal envelope modulations that carry speech information.

3. Results

Digit-triplet and IEEE-sentence SRT outcomes are displayed in the upper two panels of Figure 1. In both experiments, SRTs improved significantly as the number of channels was increased [digit triplets: $F(6, 120) = 76.38$,

$p < 0.0001$; IEEE sentences: $F(6, 120) = 58.25$, $p < 0.0001$]. In both experiments, the improvement slowed down markedly beyond around 7 channels. Bilinear fitting of SRTs on a logarithmic number-of-electrode scale was used to establish the position of the knee point. It was found to be at 7 channels regardless of speech material. The bottom panel of Figure 1 shows a scree plot of reducing normalized Eigenvalue as a function of the number of retained factors. There, bilinear fitting also showed a knee point occurring at around 7 factors.

4. Discussion

SRTs employing CI simulations with the SPIRAL vocoder showed that, even with no simulated current spread, and across two very different sets of speech material, speech intelligibility improves less steeply beyond a knee point at around 7 channels. This was also noticeable in Grange *et al.* ([3], experiment 1), where normalised, percent-correct intelligibility of IEEE sentences was presented as a function of number of activated channels and severity of current spread; even with no current spread simulated, intelligibility plateaued. The knee points in both percent-correct and SRT measures in the absence of current spread alerted us to the possibility that speech statistics might fundamentally limit the number of effective channels for speech intelligibility with CIs.

The bottom panel of Figure 1 presents the FA scree plot for speech in the absence of noise. There is a strong similarity between the scree plot and the IEEE-sentence SRTs trends, since both exhibit a knee point around 7 factors/activated electrodes. This similarity suggests that the number of effective (or independent) channels in speech intelligibility with CIs is fundamentally limited by speech statistics. As demonstrated in Grange *et al.*, the effective number of frequency channels can be further reduced by the spectral smearing caused by current spread. It should be noted that comparison is made here between FA of speech in quiet against SRTs in noise. The FA in quiet demonstrates the distribution of information across frequency in the speech, but takes no account of the robustness of this information to noise contamination. A further development of this work would therefore be to conduct FA using speech noise mixtures at appropriate SNRs.

Figure 2 shows how the factor loadings deviate from logarithmic spacing. The number of factors increases as one descends the panels. For the most part, the factors are discrete spectral bands, but the three-factor panel shows a factor that is split into two spectral peaks. Those two peaks become separate factors once a fourth factor is considered. A broader factor, centred on 900 Hz is present in most of the panels. From five to twelve factors, the factors vary significantly in their widths. In contrast, most commercial CIs analyse sound through filters that are, or approximate equal widths on a log-frequency scale. All have filters whose bandwidths increase monotonically with frequency.

The fact that speech information is grouped across frequency in factors whose widths are not logarithmically spaced begs the question of whether current CIs analyse speech signals optimally. Indeed, allocation of FA-inspired channels to an array of CI electrodes may well do a better job at transmitting speech information.

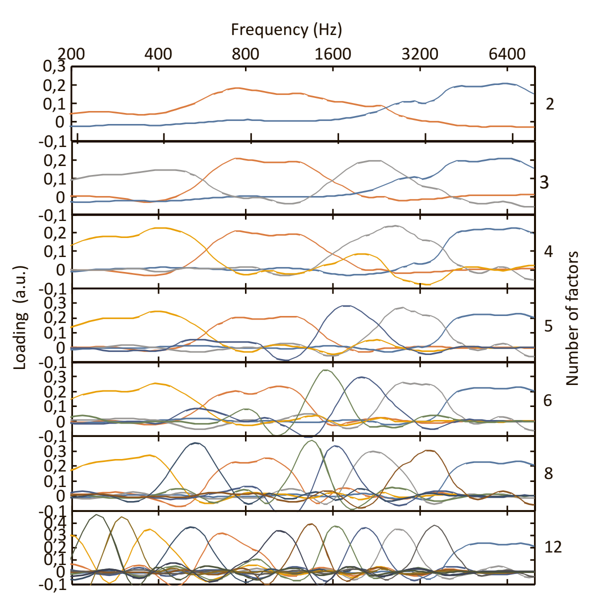

Figure 2. Outputs of the factor analysis of IEEE sentences spoken by the M.I.T. talker 'DA': factor loading curves as a function of frequency. From the top to the bottom panel, the number of factors is increased to illustrate how speech information is distributed as a function of retained factors.

Ming and Holt [5] explored the same possibility, but using a quite different method. They used a computational model of efficient auditory encoding to optimize a set of functions for the information carried by speech on the basis of the sound waveform (i.e. not limited to speech envelopes). They analysed speech into a set of six kernel functions that, once optimized, displayed band-pass characteristics reminiscent of a six-channel filterbank, but with more channels concentrated at low frequencies than in a cochleotopic map. Moreover, a vocoder based on this new filterbank produced better speech recognition than a cochleotopic one. Our FA-derived channels look rather different to those produced by Ming and Holt; the six-factor solution shows finest resolution at mid frequencies (Figure 2). Comparisons between the two designs, and the logarithmic channel distribution conventionally used in cochlear implants have yet to be made.

The information falling within these FA-inspired channels can be allocated to the electrode array in two ways. First, information from a channel can be allocated to the electrode whose place frequency is nearest to that of the channel centre frequency. This would result in a spectrally 'natural' allocation of information. However, such an FA-inspired strategy may be undermined by the effects of current spread; neighbouring narrow channels could excite closely grouped electrodes, causing information in these channels to be blended once again by current spread. A second strategy, that counters current spread, could use activated electrodes that are spaced out as much as possible along the cochlea, producing a warped frequency map designed to spatially separate the independent modulators. This strategy involves significant spectral warping, that may render the sound less natural, and consequently require more listener adaptation.

The spectrally warped FA-inspired strategy would address current spread optimally by minimising interaction between channels. From an informational point of view, such a strategy should improve intelligibility by providing more information to a CI patient's brain, and this improvement should manifest itself as steeper improvement in intelligibility with increasing numbers of channels than seen with logarithmic spacing. The trade-off between the number of channels used and counteracting the effect of current spread will still be present, however, and may lead to a number of channels for which speech intelligibility reaches a maximum. Speech FA strongly suggests that that optimum number will be lower than the 12-22 electrodes commercial CIs currently employ. The evaluation of FA-inspired strategies such as those proposed above will be the object of a follow-up study.

An important caveat is that the FA output is illustrated here for a specific voice and hence, filters inspired by such FA are optimized for that voice. The robustness of a single FA-inspired strategy can be assessed by the analysis of the effect different voices have on channel boundaries. One approach could be to establish the effect of changes in fundamental frequency and vocal tract length. Speaking style (e.g., ordinary vs. clear speech, emotional vs. neutral, varying rate of speech) and material type (e.g. connected discourse vs. short utterances) may also impact channel boundaries. The preliminary analysis of 6 different voices (three male and three female) uttering the IEEE sentences showed that for 6 to 12 channels, channel boundaries typically varied by less than ±20% in frequency terms. It is unclear at this point whether this variability may be too great for a single mapping to work for different voices.

Previous efforts made to reduce the effect of current spread, at source (e.g. Srinivasan et al [6]) or by deactivation of the least effective channels (Noble *et al.* [7]) have yielded very modest improvements. FA-inspired strategies provide a new handle to mitigate current spread, which merits careful investigation.

Acknowledgement

Work supported by the Oticon Foundation.

References

[1] P. J. Abbas, M. L. Hughes, C. J. Brown, C. A. Miller, H. South: Channel interaction in cochlear implant users evaluated using the electrically evoked compound action potential. Audiol. Neuro-Otology **9** (2004) 203–213.

[2] L. M. Friesen, R. V. Shannon, D. Baskent, X. Wang: Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. J. Acoust. Soc. Am., 110 (2001) 1150–1163.

[3] J. A. Grange, J. F. Culling, N. S. L. Harris, S. Bergfeld.: Cochlear implant simulator with independent representation of the full spiral ganglion. J. Acoust. Soc. Am. **142** (2017) EL484-EL489.

[4] K. Ueda, Y. Nakajima: An acoustic key to eight languages/dialects: Factor analyses of critical-band-filtered speech. Nature Scientific Reports **7** (2017) 42468.

[5] V. L. Ming., L. L. Holt: Efficient coding in human auditory perception. J. Acoust. Soc. Am. **126** (2009) 1312–1320.

[6] A. G. Srinivasan, M. Padilla, R. V. Shannon, D. M. Landsberger: Improving speech perception in noise with current focusing in cochlear implant users. Hear. Res. **299** (2013) 29–36.

[7] J. H. Noble, R. H. Gifford, A. J. Hedley-Williams, B. M. Dawant, R. F. Labadie: Clinical evaluation of an image-guided cochlear implant programming strategy. Audiol. Neuro-Otology **19** (2014) 400–411.