Understanding vocoded speech with two types of upward spectral shift: place-frequency mismatch in cochlear implants vs. shrinking speakers

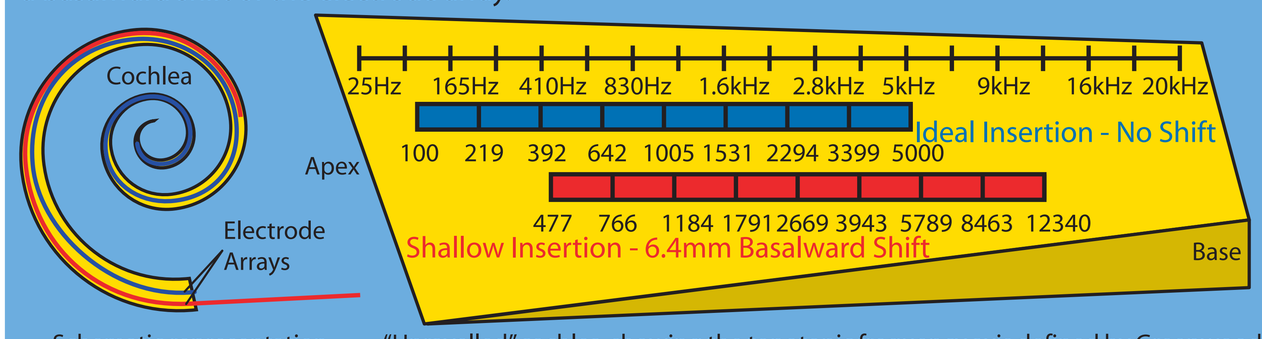
Alexis Hervais-Adelman, Annick Aubin-Pouliot, Melanie Knowles, Roy D Patterson

Centre for the Neural Basis of Hearing, Physiology, Development and Neuroscience Department, University of Cambridge, Downing Street, Cambridge, CB2 3EG United Kingdom.

1. Introduction

Cochlear implants (CIs) restore the sensation of hearing in profoundly deaf individuals by stimulating the auditory nerve with an electrode array placed in the cochlea. Due to the morphology of the cochlea, electrode arrays cannot be inserted to the apex. Thus, sounds transduced by the device stimulate regions of the cochlea that code for higher frequency signals than the source. This basalward shift induces a place-frequency mismatch, which is thought to severely reduce the intelligibility of speech signals transduced by CIs.

Noise Vocoding is a technique used to simulate CI processors (Shannon et al., 1995) for normally-hearning listeners. Upward spectral-envelope shifts can be applied to NV speech to simulate a basalward shift of the electrode array.



Schematic representation of a human cochlea, showing an ideal insertion of an electrode array and a shallow (typical) insertion.

"Unravelled" cochlea, showing the tonotopic frequency axis defined by Greenwood (1990), and two example 8-channel electrode arrays of a putative cochlear implant, which processes sound between 100-5000Hz. The ideal array delivers sound to the appropriate cochlear places, whereas the shallow one delivers sound to places coding for higher frequency than the analysed sounds.

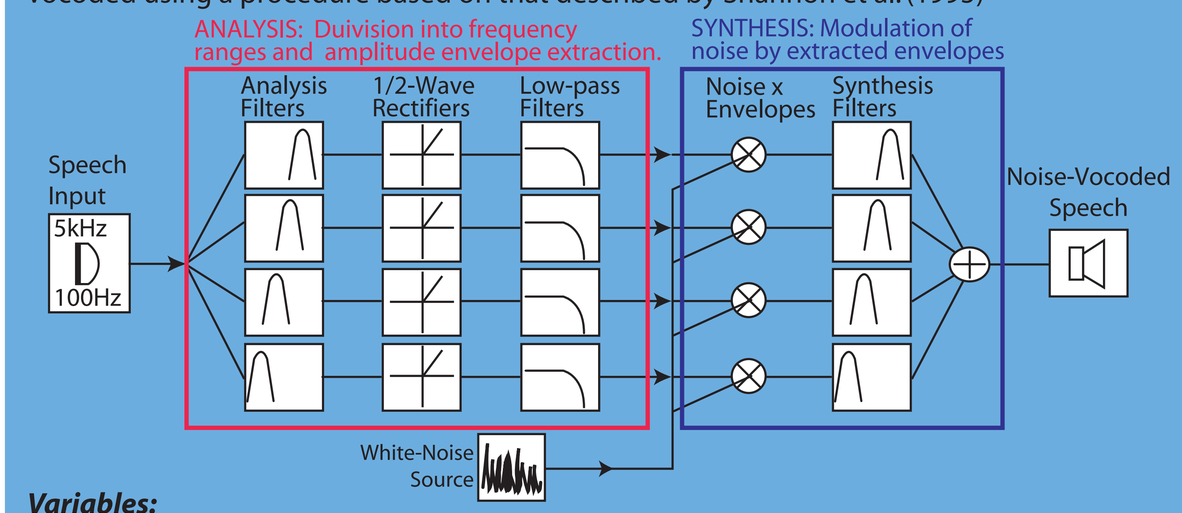
A cochlear place-frequency mismatch shifts the spectral envelope along the cochlear (tonotopic) frequency axis, which deviates from log-linear (Greenwood, 1990), distorting the spectral envelope, and key features of speech, such as formant ratios. It has been shown that upward spectral shifts of NV speech can have catastrophic consequences for intelligibility (e.g. Fu & Shannon, 1999; Rosen, Faulkner & Wilkinson, 1999).

In contrast, recent experiments have demonstrated that humans are very good at understanding speech whose spectral envelope has been shifted upwards to simulate shrinking of the speaker (e.g. Smith et al., 2005 and Ives et al., 2005). Simulating a speaker size change moves the whole spectral envelope of speech along a log(frequency) axis by a scaling factor, preserving formant ratios. and the shape of the spectral envelope.

The present work investigates whether the difficulty in comprehending spectrally-shifted noise-vocoded speech stems from the relative warping of the signal in the frequency domain, the spectral degradation imposed by noise-vocoding or a combination of these factors.

2. Method

90 consonant-vowel (CV) syllables, spoken by a native speaker of Canadian English were noise-vocoded using a procedure based on that described by Shannon et al. (1995)



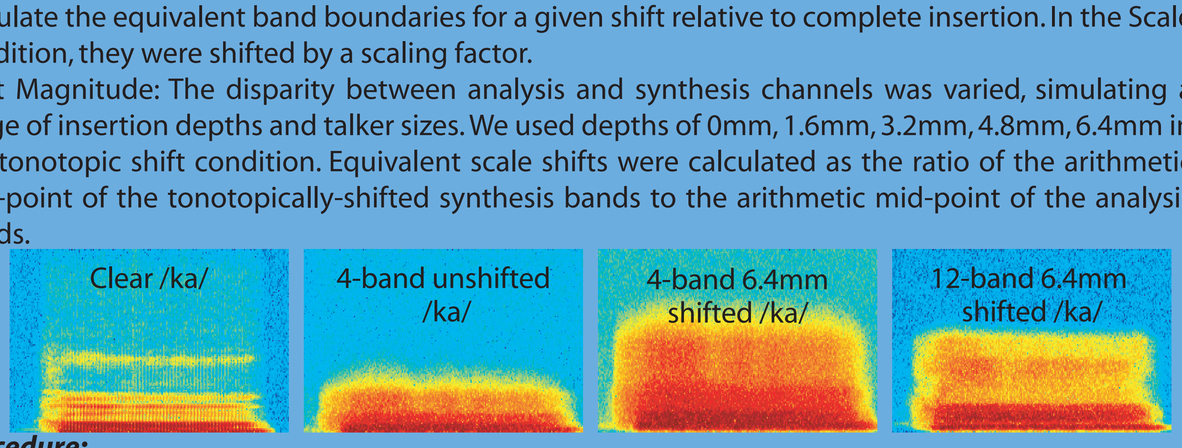
Variables:

Spectral Detail: Stimuli were vocoded with 4, 6, 8 and 12 bands. The greater the number of bands, the greater the degree of spectral detail preserved.

Shift Type: Tonotopic (simulating a shifted CI electrode array) or Scale (simulating a shrinking talker).

The spectral envelopes of the stimuli were shifted by altering the frequencies of the synthesis filters. In the tonotopic condition, synthesis bands were defined using Greenwood's equation to calculate the equivalent band boundaries for a given shift relative to complete insertion. In the Scale condition, they were shifted by a scaling factor.

Shift Magnitude: The disparity between analysis and synthesis channels was varied, simulating a range of insertion depths and talker sizes. We used depths of 0mm, 1.6mm, 3.2mm, 4.8mm, 6.4mm in the tonotopic shift condition. Equivalent scale shifts were calculated as the ratio of the arithmetic mid-point of the tonotopically-shifted synthesis bands to the arithmetic mid-point of the analysis bands.



Procedure:

8 Normally-hearing native-English speaking listeners. 90-alternative forced choice paradigm. Listeners heard 3 vocoded syllables from one "speaker" (i.e. one permutation of number of bands, shift type and shift magnitude) over headphones. They were asked to identify the third sound by selecting it on a grid containing the 90 possible CVs. An initial training period ensured familiarity with the task and response set.

40 conditions (5 shift magnitudes x 4 channels x 2 shift types)

4 trials per condition per run, 20 runs per participant.

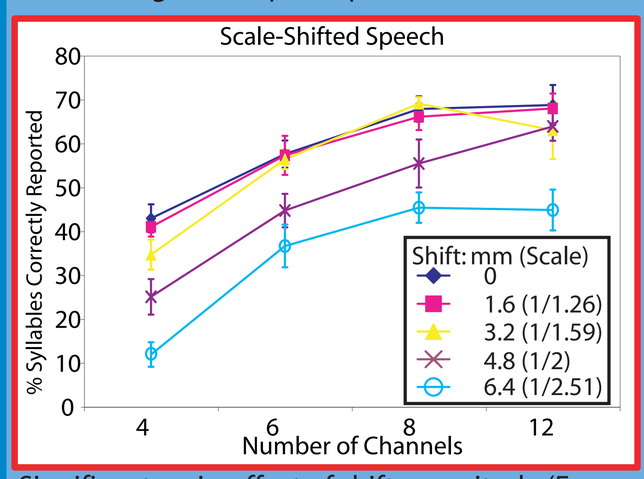
However, due to experimenter error some data were not recorded, and there were a total of 344 trials per condition for the scale-shifted stimuli and 468 per condition for the tonotopically-shifted stimuli

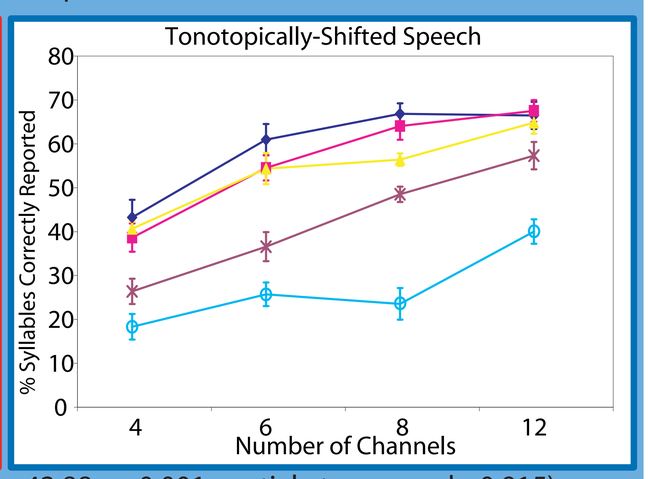
Acknowledgements

Research supported by the U.K. Medical Research Council (G0500221, G9900362).

3. Results

Data averaged over participants and trials. Error bars represent +/-1 standard error of the mean.





Significant main-effect of shift magnitude ($F_{(4,16)}$= 42.88, p<0.001, partial eta-squared= 0.915).
The greater the simulated shift, the more difficult the syllables are to correctly identify.

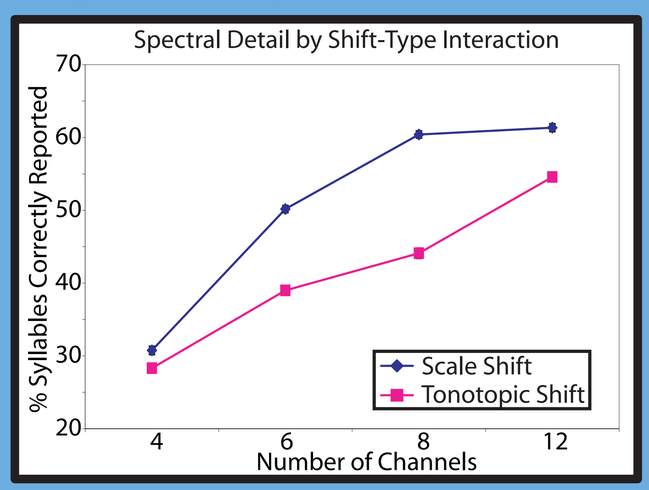
Post-hoc pairwise comparisons show that performance in the 6.4mm shift is significantly worse than any other condition, 4.8mm shift performance is signifcantly worse than no shift or 1.6mm.

Significant main-effect of spectral detail ($F_{(3,12)}$= 333.64, p<0.001, partial eta-squared= 0.988).
Decreasing spectral detail renders syllables more difficult to identify.

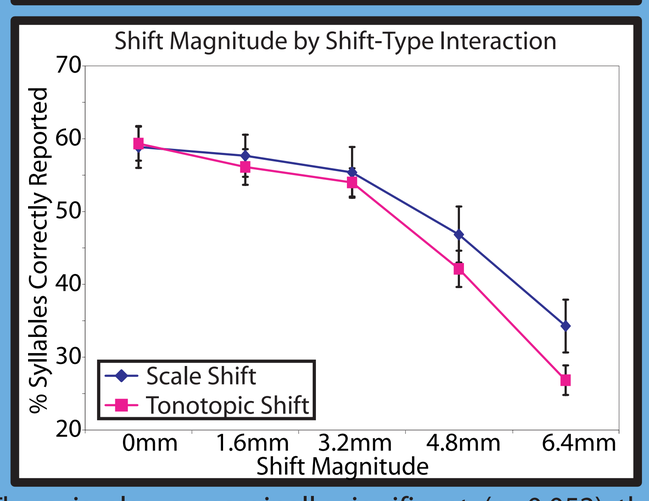
Post-hoc pairwise comparisons show that there is a significant difference in performance between all pairs of spectral detail conditions, except 8 vs 12 channels which is marginally significant (p=0.054).

Main effect of shift type is not significant (p=0.086).

Although there is no significant main effect of shift type, it is involved in significant interactions.



There is a significant interaction of spectral detail and shift type ($F_{(3,12)}$= 6.138, p= 0.009, partial eta-squared= 0.607).
The graph shows the data collapsed over all shift magnitudes, and it can be clearly seen that at intermediate levels of spectral detail, syllable-identification performance is significantly more impaired by spectral degradation in the tonotopically-shifted condition than the scale-shifted condition.



There is a significant interaction of shift magnitude and shift type ($F_{(4,16)}$= 3.057, p= 0.048, partial eta-squared= 0.433).
The graph shows the data collapsed over all levels of spectral detail, and it can be seen that syllable identification performance is affected more by spectral shift in the tonotopic than the scale shift conditions

There is also a marginally-significant (p=0.052) three-way interaction between shift type, shift magnitude and spectral detail, such that the effect of spectral detail on intelligibility is more deleterious as the magnitude of shift increases, and this effect is more pronounced in the tonotopic -shift than scale-shift condition.

4. Conclusions

The data indicate that under certain circumstances, the tonotopic shift induced by the place-frequency mismatch of CI electrode arrays at shallow insertion depths is more deleterious to speech intelligibility than are the scale shifts produced by artificially "shrinking" a speaker.

While tonotopic shifts are not encountered in natural listening environments, humans speakers of various sizes do exist, and the auditory system appears to have a mechanism to cope with this type of variability.

Replicating a finding by Fu & Shanon (1999), there is no interaction between the effect of spectral detail and shift magnitude on intelligibility, suggesting that even at low spectral resolution there are still cues that permit the normalisation of shifted speech.

Given that the auditory system is geared to handling scale shifts, if CIs could minimise the warping of the frequency axis induced by the tonotopic shift of the electrode array this could enhance intelligibility of CI speech.

Pre-shifting the spectral envelope of speech upwards, so that the disparity between analysis and synthesis filters in the cochlea occurs in a region where the tonotopic and log(frequency) axes are less divergent could help in this regard, although information about speaker size would be lost.

References

Fu, Q. J., & Shannon, R. V. (1999). Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing. *J Acoust Soc Am, 105*(3), 1889-1900.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species--29 years later. *J Acoust Soc Am, 87*(6), 2592-2605.

Ives, D. T., Smith, D. R., & Patterson, R. D. (2005). Discrimination of speaker size from syllable phrases. *J Acoust Soc Am, 118*(6), 3816-3822.

Rosen, S., Faulkner, A., & Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: implications for cochlear implants. *J Acoust Soc Am, 106*(6), 3629-3636.

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

Smith, D. R., Patterson, R. D., Turner, R., Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *J Acoust Soc Am, 117*(1), 305-318.