Comparing live to recorded speech in training the perception of spectrally shifted noise-vocoded speech

Andrew Faulkner,[a] Stuart Rosen, and Tim Green
Speech Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, United Kingdom
a.faulkner@ucl.ac.uk, stuart@phon.ucl.ac.uk, tim.green@ucl.ac.uk

Abstract: Two experimental groups were trained for 2 h with live or recorded speech that was noise-vocoded and spectrally shifted and was from the same text and talker. These two groups showed equivalent improvements in performance for vocoded and shifted sentences, and the group trained with recorded speech showed consistently greater improvements than untrained controls. Another group trained with unshifted noise-vocoded speech improved no more than untrained controls. Computer-based training thus appears at least as effective as labor-intensive live-voice training for improving the perception of spectrally shifted noise-vocoded speech, and by implication, for training of users of cochlear implants.

PACS numbers: 43.71.Es, 43.71.Ky, 43.66.Ts [QJF]
Date Received: July 25, 2012 **Date Accepted:** September 6, 2012

1. Introduction

The ability of human listeners to adapt to distortions of speech is of interest both in relation to the effectiveness of auditory prostheses and in more basic investigations of speech perception. This ability also raises the question of how such adaptation can be facilitated through training. Much work has centered on noise or tone-vocoding, which has become popular as a technique for manipulating the level of spectral detail, and because of its similarity to the processing in cochlear implants. Listeners can adapt rapidly to low spectral resolution vocoded speech (e.g., Davis *et al.*, 2005). However, a combination of vocoding with spectral shifting can be more challenging. When vocoded speech is spectrally shifted upward to simulate relatively shallow CI electrode insertions, shifts in excess of 3 mm of basilar membrane distance have large acute effects on speech perception (Dorman *et al.*, 1997; Shannon *et al.*, 1998). These effects can be markedly reduced with training, but this requires several hours, substantially longer than for vocoding alone (Faulkner *et al.*, 2006; Fu and Galvin, 2003; Nogaki *et al.*, 2007; Rosen *et al.*, 1999). Cochlear implant listeners show comparable adaptation to changes of frequency mapping (Dorman and Ketten, 2003; Fu *et al.*, 2002), and several studies indicate that explicit speech-based training can facilitate this (Fu *et al.*, 2005; Fu and Galvin, 2008).

Connected discourse tracking (CDT: DeFilippo and Scott, 1978) has been reported as effective in training normal hearing listeners to adapt over several hours to speech that is vocoded and spectrally shifted (Faulkner *et al.*, 2006; Rosen *et al.*, 1999; Smith and Faulkner, 2006). The learning shown in these studies was more than expected simply from repetition of test materials, but in the absence of control groups, the effectiveness of CDT was not absolutely established. CDT uses live voice and a dyadic interaction that mimics many features of natural communication. The trainer reads phrases from a connected text that the trainee attempts to repeat back. If the

[a]Author to whom correspondence should be addressed.

repetition is accurate, the trainer proceeds to the next phrase. If part of the response is in error, the trainer repeats the phrase, and another response is elicited. CDT encourages the use of naturalistic idioms and allows the training talker to adapt their speaking style to perceptual difficulties (Hazan and Baker, 2011). However, the use of live-voice makes CDT labor-intensive, and in both research and clinical contexts, automated training has many attractions. In clinical contexts, where the object is to improve communication, automated training also has cost advantages. In research, the use of unrepeatable live-voice compromises comparisons across conditions.

Here we have extended a method of computer-based training previously used with sentences to make it closely comparable to live-voice CDT, so allowing the importance of live-voice presentation to be assessed. This method is based on techniques designed to share some of the features of CDT (Fu *et al.*, 2005; Stacey and Summerfield, 2007). We have followed closely the implementation of Stacey and Summerfield using a closed-set task in which sentences were replaced by pre-recorded phrases from the same connected narrative text and talker as used for live-voice CDT. We therefore have two training methods that are identical in respect of the training talker and text and differ in the use of live compared to pre-recorded speech, in the use of an open response choice in CDT compared to a closed response set, and in the degree of interaction inherent to the methods. A comparison of these two methods allows us to assess whether there are advantages from the more naturalistic approach of live-voice CDT. The speech processing is similar to that used in two previous studies (Faulkner *et al.*, 2006; Stacey and Summerfield, 2007) and simulates an eight-channel cochlear implant with upward frequency shifting representing the relatively shallow electrode insertions that are found in many CI recipients (Ketten *et al.*, 1998). The choice of eight channels reflects the effective number of channels seen in CI users who perform relatively well in noise (Friesen *et al.*, 2001).

2. Method

2.1 Subjects

Four groups of 12 young normal-hearing adult native speakers of British English took part. Two groups underwent training with noise-vocoded, spectrally shifted speech. Live-voice CDT was used with one of these groups (CDT-Live group). The second (PC-Shifted) group received PC-based training with pre-recorded speech from the same text and talker. There were also two control groups. One control group received no training. A second control group (PC-Unshifted) received the same PC-based training as the PC-Shifted group except that the speech they heard during training was noise-vocoded *without* spectral shifting. This allows a test of the specificity of the training to the spectral shift aspect of the speech processing. Both control groups were exposed to shifted vocoded speech only in testing.

2.2 Speech materials

One young adult female talker of standard southern British English (SSBE) was the training talker in both pre-recorded and live-voice conditions. The training text was a graded reader for students of English (Hardcastle, 1975). Such texts are consistent in complexity and controlled in their vocabulary and syntax. The text contained 5018 words that were divided into 902 phrases of 2–11 words with a median phrase length of 5. Testing materials were from the training talker and from two additional SSBE talkers (one male, one female). These comprised IEEE sentences (IEEE, 1969), BKB sentences (Bench *et al.*, 1979), IHR sentences (MacLeod and Summerfield, 1990) and 10 vowels /æ e ɪ ɒ ʌ ɑː iː ɜː ɔː uː/ in /bVd/ words with five tokens of each vowel per talker. The BKB and IHR materials are similar in construction and were treated as equivalent. For sentence materials, specific lists were counterbalanced across conditions between subjects.

2.3 Speech processing

Real-time noise-excited vocoder processing was implemented using the ALADDIN WORKBENCH software (Hitech AB) and two Loughborough Sound Images TMS320C30 DSP

Table 1. Band cut-offs in hertz. The upper cut-off of bands 1–7 matched the lower cut-off of the band above.

Channel	1 lower	2 lower	3 lower	4 lower	5 lower	6 lower	7 lower	8 lower	8 upper
Analysis filter	100	214	378	612	947	1427	2113	3095	4500
Shifted carrier	443	705	1080	1616	2384	3482	5054	7304	10 522

cards. The eight analysis filters spanned 100–4500 Hz and were spaced at equal basilar membrane distances according to Greenwood's (1990) cochlear position map. These filters were elliptic designs with three orders per side and cut-off frequencies as shown in Table 1. An envelope was extracted from each analysis band using half-wave rectification and a 160 Hz low-pass filter (fourth order elliptic). Each band envelope was then multiplied against an independent white noise. The resulting modulated noises were passed through eight output filters and finally summed together. In the unshifted vocoder, the output filters matched the analysis filters. In the shifted vocoder, the output filters (see Table 1) had cut-off frequencies shifted upwards from the analysis filters by 6 mm on the basilar membrane according to Greenwood's map. Processed stimuli were presented to both ears at a level of approximately 70 dB SPL through Sennheiser HD540 headphones.

2.4 Procedure

An initial familiarization and pre-training test session lasted about 90 min. A total of 2 h of training was split over 2 days, with the post-training test immediately following the second hour of training. Familiarization comprised: 10 min of live CDT using unshifted vocoded speech; a 30-item vowel test using first unprocessed speech and then unshifted vocoded speech; 10 IEEE sentences and 32 BKB or 30 IHR sentences, all processed with the unshifted vocoder. The talker was varied between test materials for each subject. Testing was performed with both the shifted and unshifted vocoder for all materials. In each of the pre- and post-training test sessions, 20 IEEE sentences and 32 BKB sentences or 30 IHR sentences were presented for each talker. There was also a vowel test with 50 tokens from each talker. The order of the shifted and unshifted conditions was balanced between subjects and sessions, while the order of the talkers was randomized. Tests were blocked first by talker, then condition, and finally by speech material.

For CDT, the trainer and the trainee were in adjoining rooms. A computer prompted the trainer with each phrase. Speech from a microphone (Laryngograph Ltd. PCLX processor) was noise-vocoded and presented to the trainee over headphones. The trainer was able to hear the trainee's responses over an intercom. If any words were incorrectly repeated, the trainer repeated the phrase with a maximum of three presentations of each phrase. If the phrase was still not correctly repeated, the trainer displayed the phrase as text on a computer monitor and spoke the phrase a final time. The trainer then moved on to the next phrase. Training was run in blocks of about 10 min with breaks of less than 1 min between blocks. The total duration of CDT training over the two sessions was 2 h. The number of phrases completed during training is presented in Sec. 3.

For computer-based training, training material was also presented phrase by phrase with the phrases matching exactly those used in CDT. The trainee heard a phrase and saw a computer display showing between one and four target words (median three) and the same number of foil words. The targets were always content words and excluded proper names. Foils shared at least two phonemes with the target. The trainee selected the words s/he believed had been presented. If a foil was selected, the phrase was immediately replayed, and this process continued until all targets had been checked. At this point, the phrase was displayed orthographically and played out once more. The procedure then moved to the next phrase. As with CDT, the total duration

of training was 2 h. Training ran in 10 min blocks and subjects were able to pause briefly between these.

3. Results

Performance with unshifted vocoded speech was substantially better than in the shifted condition and was often near perfect. An ANOVA showed no effects of talker, group, or training for any of the tests nor interactions involving these factors. Hence analyses focused only on the shifted vocoded condition. Data were analyzed using mixed model ANOVA with factors of subject, group, training, and talker and Bonferroni-corrected *post hoc* comparisons.

3.1 Sentences

Pre- and post-training key word scores for shifted vocoded sentences are shown in Fig. 1. IEEE scores improved from about 15% to over 30% for the CDT-Live and PC-Shifted groups, while post-training scores were around 20% for the untrained and PC-Unshifted groups. A similar pattern was shown for BKB/IHR sentences, although scores were always higher with these simpler sentences. In each case, there were significant main effects of training, group, talker, an interaction of group and training, but no interaction of training with talker. To simplify the interpretation, difference scores between pre- and post-training sessions were then analyzed. Both sentence tasks showed a significant effect of group; (IEEE: $F[3,126] = 9.7$, $P < 0.001$; BKB/IHR: $F[3,73] = 4.1$, $P = 0.009$). For IEEE sentences only, there was a significant effect of talker ($F[2,83] = 5.37$, $P = 0.006$). The difference scores showed no interactions between group and talker, indicating that training effects were equivalent for test materials from the training talker and from the other two talkers. For IEEE sentences, all groups showed some improvement between pre-and post-training tests, but changes with training for the CDT-Live and PC-Shifted groups were significantly greater ($P < 0.05$) than for the untrained and PC-Unshifted groups. The CDT-Live and PC-Shifted groups showed statistically equivalent improvements, while the change in the PC-Unshifted group did not differ from the untrained group. For BKB/IHR sentences, again all groups improved. Here only the PC-Shifted group showed significantly larger improvements than the untrained group, and no other paired comparisons were significantly different.

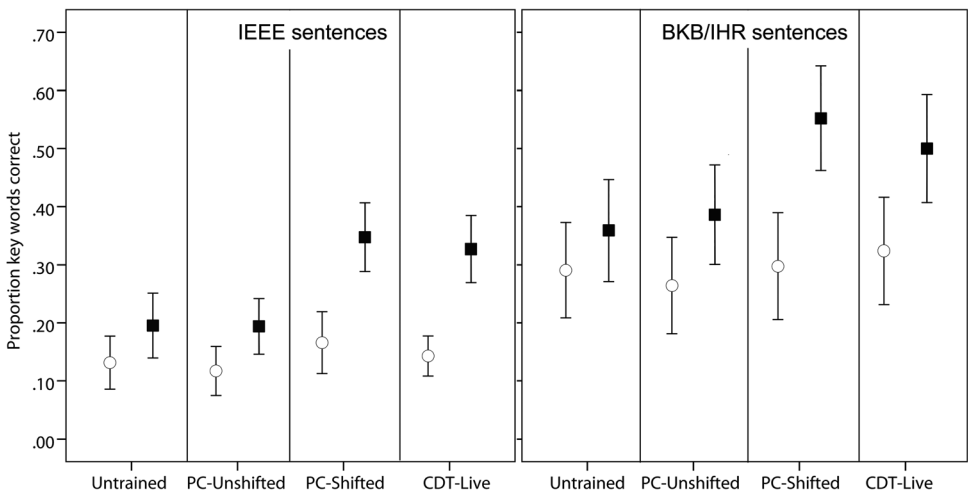

Fig. 1. Sentence scores with shifted vocoded speech before (unfilled circles) and after training (filled squares) for the four different groups. Scores are averaged over the three talkers. Error bars represent 95% confidence limits.

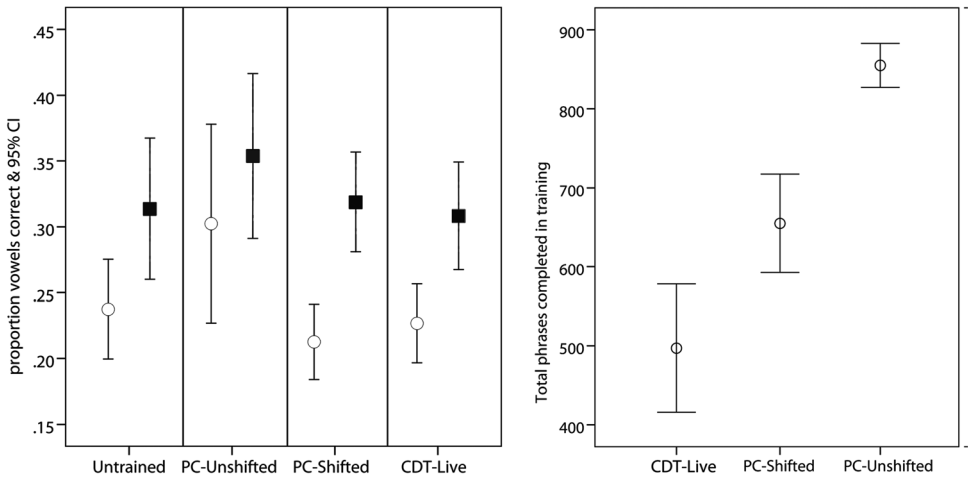

Fig. 2. Left panel (a): Vowel identification with shifted vocoded speech before (unfilled circles) and after training (filled squares) for the four groups. Right panel (b) Number of phrases completed during training by the three trained groups.

3.2 Vowels

Vowel scores for shifted vocoded speech are shown in Fig. 2(a). All four groups improved from pre- to post-training tests but never by more than 11 percentage points. An ANOVA showed an overall difference between pre- and post-training scores but no interaction with group, hence improvements were equivalent across all four groups, including the untrained group. There were significant main effects of group and test talker but no interactions involving these factors.

3.3 Rate of progress in training

The number of phrases presented to subjects in each of the trained groups is shown in Fig. 2(b). All groups differed significantly from each other. Comparing the two PC-based training groups, where procedures were identical except for spectral shifting, progress was considerably faster with unshifted vocoding. That the process was sensitive to the presence of spectral shifting demonstrates that subjects were processing the acoustic content and not performing the selection of target words purely from the choices visible to them. Comparing the two groups of those who were trained with shifted vocoded speech, the PC-Shifted group was exposed to significantly more phrases than the CDT-Live group. This may in part reflect the differing criteria of the two methods for moving on to the next phrase. While the PC-based training required the listener to identify only key words from a small closed set, the CDT method required all words to be correct, and the choice was from an unrestricted response set.

4. Discussion

For noise-vocoded speech with a frequency shift equivalent to a 6 mm basalward basilar membrane displacement, both simple and more demanding sentence materials showed significant improvements in performance that can be attributed to training and not simply to repetition of the tests. Further, for vocoded and shifted speech, computer-controlled training was as least as effective as training using live speech from the same talker and text. The computer-based method allowed listeners to be exposed to more training phrases than the live CDT method, and it is possible that this increased exposure contributed to its effectiveness. No improvement for vocoded and shifted sentence materials was evident when the training stimuli were noise-vocoded *without* spectral shifting, demonstrating learning that is specific to the shift. Further, the improvements were comparable for the training talker and the two other test

talkers; hence the information that was learned is not strongly talker-specific. This result does not, of course, argue that the use of a single talker is ideal in training for spectrally distorted speech. While Stacey and Summerfield (2007) compared single- to multi-talker training in similar conditions to the present study and found no advantage for multiple talkers, they did report an advantage of multi-talker training for a smaller 3 mm basalward shift. Clearly, multi-talker training is more readily provided with pre-recorded materials than with live voice.

The training methods used here did not lead to larger improvements in vowel identification than seen in the control conditions. It may be that 2 h of training with connected prose is not sufficient. Other studies suggest that training effects may be specific to the test materials. For example, Stacey and Summerfield (2008) found that 3 h of sentence training using a similar shifted noise-vocoder led to improved sentence recognition but not to improved identification of vowels or of consonants. A complementary finding reported by Fu *et al.* (2005) is that vowel recognition was improved by vowel and consonant phoneme training but not by sentence training.

In conclusion, computer-controlled interactive training using recordings of a connected text as training material was as effective as more labor-intensive live-voice CDT in improving the perception of sentences subjected to noise-vocoding and frequency shifting. It seems at least plausible that the same is true in the clinical training of CI users.

Acknowledgments

This work was supported by Welcome Trust Vacation Scholarship VS/07/UCL/A16 and Deafness Research UK Vacation Scholarship ref 472:UCL:AF. Thanks to Claire Watt and Kristina Gedgaudaite for running this study and to Paula Stacey for providing details of the sentence training on which this work was based.

References and links

Bench, J., Kowal., A., and Bamford, J. M. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," Br. J. Audiol. **13**, 108–112.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives; perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol. Gen. **134**, 222–241.

DeFilippo, C. L., and Scott, B. L. (**1978**). "A method for training and evaluation of the reception of on-going speech," J. Acoust. Soc. Am. **63**, 1186–1192.

Dorman, M. F., and Ketten, D. (**2003**). "Adaptation by a cochlear-implant patient to upward shifts in the frequency representation of speech," Ear Hear. **24**, 457–460.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). "Simulating the effect of cochlear-implant electrode insertion depth on speech understanding," J. Acoust. Soc. Am. **102**, 2993–2996.

Faulkner, A., Rosen, S., and Norman, C. (**2006**). "The right information may matter more than frequency-place alignment: Simulations of frequency-aligned and upward shifting cochlear implant processors for a shallow electrode array insertion," Ear Hear. **27**, 139–152.

Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (**2001**). "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**, 1150–1163.

Fu, Q. J., and Galvin, J. J. (**2003**). "The effects of short-term training for spectrally mismatched noise-band speech," J. Acoust. Soc. Am. **113**, 1065–1072.

Fu, Q. J., and Galvin, J. J. (**2008**). "Maximizing cochlear implant patients' performance with advanced speech training procedures," Hear. Res. **242**, 198–208.

Fu, Q. J., Galvin, J., Wang, X., and Nogaki, G. (**2005**). "Moderate auditory training can improve speech performance of adult cochlear implant patients," ARLO **6**, 106–111.

Fu, Q. J., Nogaki, G., and Galvin, J. J. (**2005**). "Auditory training with spectrally shifted speech: Implications for cochlear implant patient auditory rehabilitation," J. Assoc. Res. Otolaryngol. **6**, 180–189.

Fu, Q. J., Shannon, R. V., and Galvin, J. J. (**2002**). "Perceptual learning following changes in the frequency-to-electrode assignment with the Nucleus-22 cochlear implant," J. Acoust. Soc. Am. **112**, 1664–1674.

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.
Hardcastle, M. (**1975**). *Money for Sale* (Heineman Educational Books Ltd., Portsmouth, NH).
Hazan, V., and Baker, R. (**2011**). "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," J. Acoust. Soc. Am. **130**, 2139–2152.
IEEE. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **AU17**, 225–246.
Ketten, D. R., Skinner, M. W., Wang, G., Vannier, M. W., Gates, G. A., and Neely, J. G. (**1998**). "In vivo measures of cochlear length and insertion depth of nucleus cochlear implant electrode arrays," Ann. Otol. Rhinol. Laryngol. Suppl. **175**, 1–16.
MacLeod, A., and Summerfield, Q. (**1990**). "A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use," Br. J. Audiol. **24**, 29–43.
Nogaki, G., Fu, Q. J., and Galvin, J. J. (**2007**). "Effect of training rate on recognition of spectrally shifted speech," Ear Hear. **28**, 132–140.
Rosen, S., Faulkner, A., and Wilkinson, L. (**1999**). "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants," J. Acoust. Soc. Am. **106**, 3629–3636.
Shannon, R. V., Zeng, F. G., and Wygonski, J. (**1998**). "Speech recognition with altered spectral distribution of envelope cues," J. Acoust. Soc. Am. **104**, 2467–2476.
Smith, M. W., and Faulkner, A. (**2006**). "Perceptual adaptation by normally hearing listeners to a simulated 'hole' in hearing," J. Acoust. Soc. Am. **120**, 4019–4030.
Stacey, P. C., and Summerfield, A. Q. (**2007**). "Effectiveness of computer-based auditory training in improving the perception of noise-vocoded speech," J. Acoust. Soc. Am. **121**, 2923–2935.
Stacey, P. C., and Summerfield, A. Q. (**2008**). "Comparison of word-, sentence-, and phoneme-based training strategies in improving the perception of spectrally distorted speech," J. Speech Lang. Hear. Res. **51**, 526–538.