Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Investigating the perception of noise-vocoded speech - an individual differences approach

Carolyn McGettigan[a], Stuart Rosen[b] and Sophie Scott[a]

[a]University College London, Institute of Cognitive Neuroscience, 17 Queen Square, WC1N 3AR London, UK
[b]UCL, Wolfson House, 4, Stephenson Way, NW1 2HE London, UK
c.mcgettigan@ucl.ac.uk

We used a cochlear implant simulation (noise-vocoded speech) to investigate speech recognition and perceptual learning in hearing adult speakers of English. In two separate sessions (1-2 weeks apart), 28 listeners were tested on recognition of noise-vocoded sentences, words, consonants and vowels. There was evidence of significant perceptual learning that survived until Session 2 for all tasks. An individual differences analysis of Session 1 data suggested two independently-varying 'levels' of processing during the initial perception of the distorted speech stimuli - a 'top-down' listening mode making use of contextual and lexical information, and a 'bottom-up' mode focused on acoustic-phonetic discriminations. By Session 2, a more general listening mode emerged, reflecting consolidation of basic sound-to-representation mappings. Information Transfer analyses of consonant and vowel data suggested that better speech recognition may be achieved through more efficient use of preserved cues to duration and voicing. We conclude that training regimes involving directed attention to specific features, such as vowel length, may help to improve performance with noise-vocoded speech.

1 Introduction

The recipient of a cochlear implant is faced with the challenge of making sense of a new sound world. This process of adaptation, or *perceptual learning* can take a long time, with widely varying levels of success. Acoustic simulations of cochlear implants (e.g. noise-vocoding) have been used with hearing participants to model learning after implantation [1, 2, 3, 4]. Many of these studies employed group designs to identify successful training regimes for learning. In the current study, we avoid the use of explicit training routines and complex designs in order to describe basic adaptation to a range of noise-vocoded speech stimuli. Rather than adopt a group design, we harness individual differences to explore perceptual trends, and how these change over time.

Noise-vocoding produces speech with degraded spectral detail, by replacing the original signal by a variable number of amplitude-modulated noise bands. This degraded speech sounds like a noisy whisper, and the addition of further bands contributes to increased intelligibility. An early study employing this transformation [5] tested recognition of sentences, vowels and consonants vocoded to 1, 2, 3 and 4 bands - that study showed differences in overall performance on the different tasks, but there was no description of the relationships between them. More recently, experiments using vocoded stimuli have looked for transfer of perceptual learning across stimulus categories in paradigms where the listeners are trained on one linguistic category and tested on one or more others [1, 2, 3, 4]. Such a paradigm complicates the interpretation of cross-task relationships in the data. However, extracting patterns of covariance can offer an extra insight into the underlying perceptual processes. For example, close correlation of speech recognition at segment, word and sentence level may indicate a unified acoustic strategy from the listener, whereas statistical independence of sentence stimuli from words and segments may reflect considerable importance for top-down processing strategies.

There has been some debate about how best to measure perceptual learning of speech. One approach compares speech recognition scores in participants who had received training with those who have not [1, 2]. Other authors have included 'pre-test' measures of performance to ensure that group differences after training do not merely reflect selection effects [4]. However, these studies used a fixed level of spectral degradation for all stimuli, which introduces the risk of floor or ceiling effects. An alternative method was presented in a meta-analysis in which data sets from several cochlear implant simulation experiments were modelled with sigmoidal curves describing performance against the number of bands [6]. Comparison of the positions of these curves reflected the relative difficulty of the materials used in each experiment, but the authors found curve slope to be uninformative. A similar approach could be used to quantify individual differences in performance and changes associated with perceptual learning, within one experiment.

This experiment tests recognition of noise-vocoded sentences, words and segments at a range of degradation levels in a single group of hearing adults. Individual measures of curve position and slope are used to assess the inter-relationship of tasks; it is predicted that all the tasks will correlate significantly, but that the sentences and words tasks will covary strongly due to the effects of top-down processing in the presence of a lack of acoustic clarity. Two sentence corpora of differing overall complexity and predictability are included to explore these top-down effects. Listeners are tested on two separate occasions at least 7 days apart, and significant learning is predicted for all tasks. In contrast to previous authors' conclusions regarding slope, we hypothesize that improved performance will be associated with both a leftward shift and a steepening of the performance function.

2 Method

2.1 Participants

Participants were 28 monolingual speakers of British English (aged 18-40, 12 male), with no language or hearing problems. All were recruited from the UCL Department of Psychology Subject Pool, and were naïve to noise-vocoded speech.

2.2 Materials

Listeners were tested on perception of 5 different stimulus types, all vocoded with 1, 2, 4, 8, 16 and 32 bands. The items were also available in undistorted form. The vocoding routine followed the general scheme described in [5], with analysis and output filters between 100-5000Hz and envelope extraction via half-wave rectification and low-pass filtering at 400Hz.

Simple Sentences. One-hundred-and-forty items from the BKB sentence corpus [7], each with three keywords (e.g. The *clown* had a *funny face*).
Low Predictability Sentences. One-hundred-and-forty items from the IEEE sentence corpus [8], each with five keywords (e.g. The *birch canoe slid* on the *smooth planks*).
Single Words. One-hundred-and-forty items from the phonemically-balanced Boothroyd AB lists [9] (e.g. *gas*, *mice*, *whip*).
Consonants. Seventeen consonants: b, d, f, g, ʤ, k, l, m, n, p, s, ʃ, t, v, w, j, z. One token of each consonant was recorded in the context */aː/-C-/aː/*, where *C* is a consonant e.g. *apa*, *aga*, *ala*.
Vowels. Seventeen vowels: æ, eɪ, ɑː, ɛː, iː, iə, e, ɪ, aɪ, ɜ, ɒ, əʊ, uː, ɔː, aʊ, ɔɪ, ʌ. One token of each vowel was recorded in the context */b/-V-/d/*, where *V* is the vowel e.g. *bad*, *beard*, *boyed*.

2.3 Design and Procedure

The listeners made two visits to the lab, separated by 7-15 days ($N = 27$: $M = 10.44$ days, $SD = 2.69$), with the exception of one participant who could only return after 78 days. All stimulus presentation routines were programmed and run in MATLAB v7.1 (The Mathworks, Inc., Natick, MA).

Sentences and Words. Each session featured 70 items, with 10 at each 'distortion level'. Half of each item list was labelled as Set A and the other half as Set B. Fourteen participants received Set A items in Session 1, while the remainder received Set B items in Session 1. Within-session, a pseudorandomization routine ensured that the 70 items (i.e. their linguistic content) were completely randomized across the task, but that within each chronological block of 7 sentences there was an example from each distortion level.
Consonants and Vowels. The Consonants and Vowels were tested separately. Each of the tokens was repeated at all of the seven distortion levels, and the whole list of items was fully randomized. Exposure to the distortion levels was not chronologically constrained.

In each session, the tasks were administered in the order: BKB sentences, IEEE sentences, Words, Consonants, Vowels. All test materials were presented over Sennheiser HD25-SP headphones in a quiet room, at a fixed volume setting. The Sentences and Words tasks were open-set recognition tasks. Each stimulus was played once and the participant gave typed report of the item content. Responses were self-timed. The Consonants and Vowels tasks each adopted a 17-alternative forced-choice paradigm. The response choices were presented on a printed sheet which remained in view for the duration of the task. In these two tasks, listeners were encouraged not to leave any gaps, even when they were completely unsure of the answer.

3 Results

This section falls into two parts. In the first, psychometric performance functions are fitted to each individual's performance, and individual differences analyses of curve position and slope used to characterize group performance in the two sessions. The second part uses Information Transfer analyses to unpack the perception of consonants and vowels and relate this to recognition of sentences and words.

3.1 Psychometric performance curves

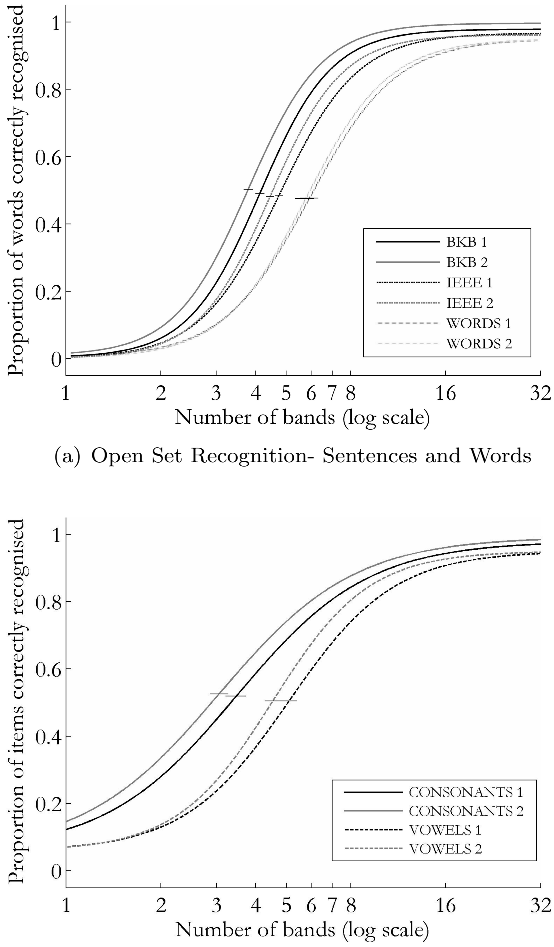
Logistic curve-fitting on each individual data set was carried out using the psignifit software package [10]. For superior fits, the distortion levels (number of bands) were converted into their log_{10} equivalents. Data from undistorted stimuli were not included. The equation used for fitting was:

$$(f(x:\alpha,\beta,\gamma,\lambda)) = \gamma + \frac{1-\gamma-\lambda}{1+e^{-(x/\alpha)^{\beta}}}$$

(1)

In the output of the fitting procedure, the α parameter corresponds to the curve's displacement along the abscissa for 50% of maximum performance), and β is inversely proportional to the curve steepness. These two parameters were extracted from each fitted curve for use in subsequent analyses. The parameter γ corresponds to the base rate of performance (or 'guessing rate'), while λ reflects the 'lapse rate' i.e. a lowering of the upper asymptote to allow for errors unrelated to the stimulus level. The software takes a constrained maximum-likelihood approach to fitting, where all four variables are free to vary, but where, in this case, γ and λ are constrained between 0.00 and 0.05. For the forced-choice tasks (Consonants and Vowels), the γ parameter was set to 1/17.

Figure 1 shows a plot of the group performance functions for the open-set (1(a)) and closed-set (1(b)) tasks in each Session. The raw scores indicated an overall decrease in α scores between Session 1 and 2, with a weaker trend in the same direction for β. A repeated-measures ANOVA analysis was run on the α scores with Session as the within-subjects factor and Task as a between-subjects factor. A second between-subjects factor, Version (which coded the order of presentation of the item sets) was also included, but no results involving this factor are reported here. There was a significant effect of Session ($F(1, 26) = 35.094$, $p = .000$, $\eta^2 = .574$, power = 1.000), a significant effect of Task ($F(4, 104) = 117.18$, $p = .000$, $\eta^2 = .818$, power = 1.000), and a non-significant interaction of these two factors ($F<1$), indicating that the degree of improvement was not significantly different across tasks. The forced-choice nature of the Consonants and Vowels tasks clearly has an effect on their slopes. For this reason, analysis of β scores was peformed in two separate ANOVAs. The first included β values from the open-set recognition tasks (BKB, IEEE and Words). This found non-significant effects of Session ($F(1, 26) = 2.80$, $p = .106$, $\eta^2 = .097$, power = .364) but a significant effect of Task (Wilks' Lambda $F(2, 25) = 3.54$, $p = .044$, $\eta^2 = .220$, power = .604). The interaction between Task and Session was non-significant ($F<1$). The corresponding ANOVA on



(a) Open Set Recognition- Sentences and Words

(b) Closed Set Recognition- Consonants and Vowels

Figure 1: Logistic curves describing group performance on the speech recognition tasks (error bars show 95% confidence limits around α).

slope parameters from the Consonants and Vowels tasks gave a non-significant effect of Session (F<1) and a non-significant interaction of Session and Task (F<1), but a significant effect of Task ($F(1, 26) = 6.00$, $p = .017$, $\eta^2 = .188$, power = .655).

There was evidence of several significant relationships across tasks for the α scores, but not between the β values. Table 1(a) shows the one-tailed Pearson's correlation matrix for α scores in Session 1. These show close intercorrelation of the Sentences and Words tasks on one hand, and the Consonants, Vowels and Words tasks on the other. A common factor analysis was run on the threshold data, with maximum likelihood extraction and varimax rotation. The rotated factor matrix is shown in Table 2(a), for those factors producing eigenvalues above 1. Two components were extracted. In the rotated matrix, the first component accounted for 22.60% of the variance, while the second component accounted for 19.21%.

The pattern of correlations for α scores in Session 2 no longer fitted the processing framework suggested by the Session 1 data (see Table 1(b)), with the Words task now somewhat separate from the others. A Common Factor Analysis (maximum likelihood extraction) was run on the data, with varimax rotation. This converged on two components - see Table 2(b). In this analysis, Factor 1 accounted for 24.41% of the variance, where Factor 2 accounted for a further 20.38%.

Table 1: Cross-task correlations - α scores.

(a) Session 1

	BKB	IEEE	Words	Cons.	Vowels
BKB	1.00	.356*	.259†	.003	-.100
IEEE		1.00	.323*	.069	-.056
Words			1.00	.417*	.331*
Cons.				1.00	.302†
Vowels					1.00

(b) Session 2

	BKB	IEEE	Words	Cons.	Vowels
BKB	1.00	.277†	.333*	.299†	.236
IEEE		1.00	-.025	.393*	.296†
Words			1.00	.015	.057
Cons.				1.00	.317†
Vowels					1.00

† =p<.10, ∗ =p<.05

Table 2: Factor analysis - α scores.

(a) Session 1

	Factor 1	Factor 2
BKB		.605
IEEE		.593
Words	.705	.469
Consonants	.558	
Vowels	.562	

(b) Session 2

	Factor 1	Factor 2
BKB	.520	.344
IEEE	.545	
Words		.946
Consonants	.642	
Vowels	.491	

showing factor loadings greater than 0.3

Table 3 shows one-tailed Pearson's correlations between α and β scores in each session, indicating that lower thresholds were generally associated with steeper performance functions. Taking a decrease in α or β to reflect an improvement in performance, the results also indicate that those listeners who exhibited the highest thresholds (i.e. most rightward curves) and shallowest slopes in Session 1 were those who showed most improvement by Session 2 (Table 4). This has been observed by other authors [3].

3.2 Information Transfer analyses

Figure 2 shows group performance curves for a selection of individual consonants in Session 1 and demonstrates that the vocoding routine affects individual speech seg-

Table 3: Correlations between α and β scores.

(a) Session 1

		β				
		BKB	IEEE	Words	Cons.	Vowels
	BKB	.259†	-.203	.045	.153	-.150
	IEEE	-.095	.096	.275†	.108	-.122
α	Words	-.164	.127	.499**	-.277†	.226
	Cons.	-.212	.136	.482**	-.169	.270†
	Vowels	-.011	-.237	.444**	-.031	.453**

(b) Session 2

		β				
		BKB	IEEE	Words	Cons.	Vowels
	BKB	-.257†	.110	-.094	-.046	.101
	IEEE	.450**	-.140	-.187	-.109	.271†
α	Words	-.076	.002	.492**	-.156	.397*
	Cons.	.065	.057	-.280†	.438**	.160
	Vowels	.140	.033	-.129	.025	.364*

$\dagger$ $=p<.10$ $* = p<.05$, $** = p<.01$

Table 4: Relationship between Session 1 performance and improvement by Session 2.

	α	β
BKB	.485**	.802***
IEEE	.751***	.726***
Words	.677***	.736***
Consonants	.550**	.863**
Vowels	.686***	.622**

$** = p<.01$, $*** = p<.001$; one-tailed.

ments differently. The forced-choice nature of the segment recognition tasks meant that the data could be arranged into confusion matrices for use in an Information Transfer analysis. This enabled description of consonant and vowel recognition in terms of the transmission of their component phonetic features. Only complete data sets, i.e those without omissions or off-list responses, were included.

Analyses were run in FIX (Feature Information XFer, UCL, UK), for each session and task separately. In the analyses, the 17 consonants were coded for the features Voicing, Place and Manner. The vowels were coded for Height, Backness, Roundedness, Length and Mono/ Diphthong status. Each analysis contained data from 14 participants. Table 5 shows the mean proportion of available information transferred for each feature.

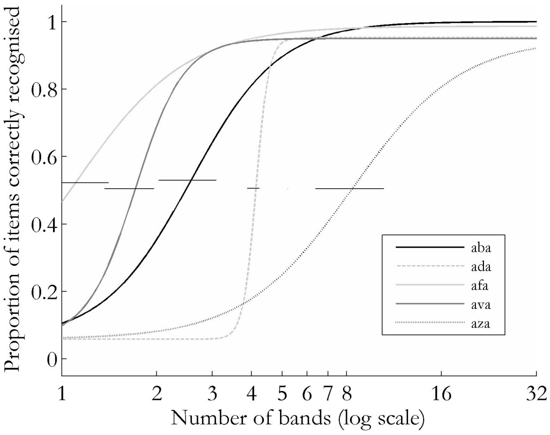


Figure 2: Group performance curves for individual consonants (error bars show 95% confidence limits around α).

Table 5: Results of Information Transfer analyses (proportion scores)

(a) Consonants

	Session 1	Session 2
Voicing	.498 (.089)	.604 (.111)
Manner	.646 (.045)	.723 (.064)
Place	.483 (.042)	.512 (.039)

(b) Vowels

	Session 1	Session 2
Height	.456 (.048)	.493 (.039)
Backness	.392 (.050)	.424 (.040)
Roundedness	.379 (.047)	.422 (.063)
Length	.632 (.183)	.707 (.181)
Diphthong	.318 (.065)	.345 (.054)

standard deviations given in brackets

The Information Transfer scores for Voicing, Place and Manner in each Session were entered as predictors in linear regression analyses on the α scores for the five overall tasks. In Session 1, a significant model with Place and Voicing as predictors offered the best account of Consonant recognition ($R^2_{adj.} = .750$; $F(2, 11) = 21.71$, $p = .001$) . Performance on the Vowels task was best predicted by Voicing ($R^2_{adj.} = .295$; $F(1, 12) = 6.44$, $p = .026$). In Session 2, Manner and Place predicted α scores on the Consonants task ($R^2_{adj.} = .580$; $F(1, 12) = 9.98$, $p = .003$), while Manner scores predicted α scores on the IEEE sentences ($R^2_{adj.} = .270$; $F(1, 12) = 5.80$, $p = .033$). A similar set of regressions was run with the five vowel features as predictors. In Session 1, a significant model featured Height as the sole predictor of α scores on the Vowels task ($R^2_{adj.} = .743$; $F(1, 12) = 38.68$, $p = .000$). In Session 2, a significant model with Height and Length ($R^2_{adj.} = .772$; $F(1, 12) = 28.51$, $p = .000$) gave the best prediction of α scores on the Vowels task, while a model with Length emerged as a significant predictor of performance on the BKB sentences ($R^2_{adj.} = .308$; $F(1, 12) = 6.80$, $p = .023$).

4 Discussion

The current data set supports the projected hypotheses. There was evidence for long-term perceptual learning of noise-vocoded sentences, words and segments. Using individual differences as the starting point for analyses, we identified a pattern of commonalities amongst the tasks, which changed with learning. Analyses of confusion data revealed predictive roles for specific phonetic features in the perception of noise-vocoded stimuli.

Factor analysis of Session 1 α scores showed two similarly-weighted, orthogonal factors, with sentences and words loading on one factor, and words and segments loading on the other. This suggests two independent modes of listening: a 'top-down' mode making use of lexical, syntactic and semantic information to generate hypotheses about stimulus identity, and a 'bottom-up' mode concerned with acoustic-phonetic discriminations. By the second session, when performance had improved, all tasks but one - Words - patterned together. It appears that once the initial learning of sound-to-representation mappings has taken place, the listener can begin to approach most stimulus types in a similar way. The plot in Figure 1(a) shows that the Words task was the most difficult of the open-set tasks in both sessions, and showed the least improvement overall by Session 2 (although a Task x Session interaction was not borne out significantly). Within the open-set tasks, the overall amount of exposure to vocoded material across seventy sentences is much greater than for seventy monosyllabic words. However, a previous study showed that, even when matched for number of words of exposure, learning is still slower for noise-vocoded words than for sentences [2]. These authors interpret such findings in terms of the relative richness of the "teaching signal" that assists learning. In the current experiment, the listener could draw upon many more sources of knowledge against which to test hypotheses for sentence recognition than they could for isolated words. Furthermore, the segment recognition tasks provided a learning framework through their forced-choice design.

This study did not directly test the finding in [6], that perceptual data for different categories of vocoded stimuli could be fitted equally well with curves of a fixed slope. However, we aimed to challenge the suggestion that slope may be uninformative. The current data set indicated that both α and β parameters decrease with perceptual learning, thus associating a leftward shift and steepening of the performance function with improved performance. However, the results were equivocal for the β parameter, as the effect of Session did not reach significance in the main ANOVA analyses. Similarly, while there was an indication of an association between the steepness of the slope and the ease of the task for open-set materials, it was also clear that overall task structure (open- versus closed-set) had a considerable effect on slope values. In conclusion, rigorous analysis of β values cannot add to the interpretation based on the α parameter.

The use of Information Transfer analyses produced findings unattainable from basic recognition scores. We identified significant roles for voicing and vowel length information in recognizing noise-vocoded stimuli. Both of these properties were well represented at low spectral resolutions in the current stimuli - in particular, vowel length information was fully present even in 1-band stimuli. However, Table 5 shows that listeners' accuracy on these features was much less than 100% in both sessions. We suggest that targeted training of discriminations based on these properties may assist in the mapping of sounds to representation in the early stages of perceptual adaptation to vocoded stimuli.

5 Conclusions

In sum, we have shown that harnessing individual differences in participants' performance can yield rich data sets with which to advance our current understanding of perceptual learning of speech.

References

[1] M. Davis, I. Johnsrude, A. Hervais-Adelman, K. Taylor, C. McGettigan, "Lexical information drives perceptual learning of noise-vocoded speech," *Journal of Experimental Psychology: General 4*(2), 254-264 (2005)

[2] A. Hervais-Adelman, M. Davis, I. Johnsrude, R. Carlyon, "Perceptual learning of noise vocoded words: Effects of feedback and lexicality," *Journal of Experimental Psychology - Human Perception and Performance 34*(2), 460-474 (2008)

[3] P. Stacey, A. Summerfield, "Effectiveness of computer-based auditory training in improving the perception of noise-vocoded speech," *Journal of the Acoustical Society of America 121*, 2923-2935 (2007)

[4] J. Loebach, D. Pisoni, "Perceptual learning of spectrally degraded speech and environmental sounds," *Journal of the Acoustical Society of America 123*, 1126-1139 (2008)

[5] R. Shannon, F. Zeng, V. Karnath, J. Wygonski, M. Ekelid, "Speech recognition with primarily temporal cues," *Science, 270*(5234), 303-304 (1995)

[6] R. Shannon, Q. Fu, J. Galvin, "The number of spectral channels required for speech recognition depends on the difficulty of the listening situation," *Journal of the Acoustical Society of America 104*, 2467-2476 (2004)

[7] J. Bench, A. Kowal, J. Banford, "The BKB (Banford-Kowal-Bench) sentence lists for partially-hearing children," *British Journal of Audiology 13*(3), 108-112 (1979)

[8] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics 17*(3), 3998-4006 (1969)

[9] A. Boothroyd, "Developments in speech audiometry," *Sound 2*, 3-10 (1968)

[10] F. Wichmann, N. Hill, "The psychometric function: I. Fitting, sampling, and goodness of fit," *Perception & Psychophysics 63*(8), 1314-1329 (2001)