Publication details: *Proceedings of the Royal Society of London, Series B: Biological Sciences, 2011, 278, 1595-1600.*

The intelligibility of noise-vocoded speech: Spectral information available from across-channel comparison of amplitude envelopes

Brian Roberts[1*], Robert J. Summers[1], and Peter J. Bailey[2]

1. Psychology, School of Life and Health Sciences, Aston University, Birmingham B4 7ET, UK.

2. Department of Psychology, University of York, Heslington, York YO10 5DD, UK.

Running title: Spectral cues in noise-vocoded speech

[*] **Correspondence to**: Professor B. Roberts at Psychology, School of Life and Health Sciences, Aston University, Birmingham, B4 7ET, UK

Email: b.roberts@aston.ac.uk

Summary

Noise-vocoded (NV) speech is often regarded as conveying phonetic information primarily through temporal-envelope cues rather than spectral cues. However, listeners may infer the formant frequencies in the vocal-tract output – a key source of phonetic detail – from across-band differences in amplitude when speech is processed through a small number of channels. The potential utility of this spectral information was assessed for NV speech created by filtering sentences into six frequency bands, and using the amplitude envelope of each band (≤30 Hz) to modulate a matched noise-band carrier (N). Bands were paired, corresponding to F1 (≈N1+N2), F2 (≈N3+N4), and the higher formants (F3′≈N5+N6), such that the frequency contour of each formant was implied by variations in relative amplitude between bands within the corresponding pair. Three-formant analogues (F0=150 Hz) of the NV stimuli were synthesised using frame-by-frame reconstruction of the frequency and amplitude of each formant. These analogues were less intelligible than the NV stimuli or analogues created using contours extracted from spectrograms of the original sentences, but more intelligible than when the frequency contours were replaced with constant (mean) values. Across-band comparisons of amplitude envelopes in NV speech can provide phonetically important information about the frequency contours of the underlying formants.

Keywords: noise-vocoded speech, spectral cues, formant frequencies, intelligibility.

1. INTRODUCTION

Speech is highly redundant and so it can remain intelligible even after substantial distortion or simplification of the signal. A commonly used simplification is vocoding, which involves filtering speech into one or more frequency bands, using the amplitude envelope of each band to modulate a carrier shaped by the corresponding filter, and reconstructing the simplified signal by summing the modulated carrier bands. The technique was originally devised by Dudley (1939) for speech transmission through telecommunications systems, particularly for encrypted communications, and has since been used widely for voice processing in popular music. Shannon et al. (1995) first introduced noise vocoding, in which the carrier for each channel is filtered Gaussian noise. Their study demonstrated that the intelligibility of noise-vocoded speech can be high when only three or four channels are used, at least when all stimuli are derived from the speech of a single talker. Dorman et al. (1997) obtained comparable results with sine-vocoded speech, a closely related stimulus consisting of a set of amplitude modulated sinusoids instead of noise bands. Vocoding has since become a standard research tool for simulating listening to speech through a cochlear implant; many contemporary studies use noise-vocoded speech (e.g., Li and Loizou 2009; Loebach et al. 2009; Chatterjee et al. 2010; Eisner et al. 2010) or sine-vocoded speech (e.g., Chen and Loizou 2010; Hopkins and Moore 2010) for this purpose.

Interpreting the results of perceptual experiments using vocoded speech requires an understanding of the nature of, and weight attached to, sources of phonetic information in the signal. Processing speech through a noise vocoder with a small number of channels implies a considerable loss of spectral information. Hence, this type of stimulus is often regarded as conveying phonetic information primarily through temporal-envelope cues rather than spectral cues (Shannon et al. 1995; Nittrouer et al. 2009). In this conception, the intelligibility of noise-vocoded speech depends mainly on the within-channel analysis of low-rate changes in amplitude over time; an account of the types of linguistic information potentially available from temporal-envelope cues has been provided by Rosen (1992). Other studies of noise- and sine-vocoded speech have tended to characterise the relative contributions of spectral and temporal cues to intelligibility in terms of the effects of varying the number of channels and the low-pass envelope cut-off, respectively, and their trade-off (e.g., Xu et al. 2005; Xu and Zheng 2007; Xu and Pfingst 2008). What is often overlooked in this characterisation is the spectral information that can potentially be retrieved through comparing the levels of adjacent channels (Dorman et al. 1997; Loizou et al. 1998). In particular, changes in relative amplitude across channels over time can potentially carry information about the underlying frequency contours of the spectral prominences in the signal, and this derived spectral information may contribute more (and temporal-envelope information perhaps less) to the intelligibility of noise- and sine-vocoded speech than is commonly supposed.

Spectral prominences in speech – called *formants* – are perceptually important, because they arise as a result of resonances in the air-filled cavities of the talker's vocal tract. Variation in the centre frequency of a formant is an inevitable consequence of change in the size of its associated cavity as the vocal-tract articulators – particularly the tongue, lips, and jaw – are moved by the talker. Thus, knowledge of formant frequencies and their change over time is likely to be of considerable benefit to listeners, as it provides salient information about the configuration and kinematics of the talker's vocal tract. The experiment reported here demonstrates that the formant-frequency contours implied by variations in relative amplitude between adjacent spectral bands can be extracted from noise-vocoded signals and can support intelligibility in synthetic-formant analogues of speech.

2. METHODS

(a) Participants

Twenty listeners (10 males) took part; their mean age was 23.2 years (range = 19.2 – 54.7). All listeners were native speakers of British English, naïve to the purpose of the experiment, and had audiometric thresholds better than 20 dB hearing level at 0.5, 1, 2, and 4 kHz. Each listener gave written consent to participate in the experiment, which was approved by the Aston University Ethics Committee.

(b) Stimuli and conditions

All stimuli were derived from 24 BKB sentences (Bench et al. 1979), spoken by a British male talker of Received Pronunciation English and low-pass filtered at 5 kHz. There were four conditions in the experiment, corresponding to the four speech analogues described below. Figure 1 shows the spectrogram of an example sentence and of the four analogues derived from it. For each listener, the sentences were divided equally across conditions (i.e., six per condition) using an allocation that was counterbalanced by rotation across each set of four listeners tested. Each sentence group was balanced so as to contain 95 or 96 phonemes in total. Examples of the stimuli are available in the electronic supplementary material.

Figure 1 near here

Noise-vocoded (NV) stimuli were created from the original sentences using Praat software (Boersma & Weenink 2008). The speech was first filtered, using a 16th-order Butterworth filter (96 dB/octave roll-off), into six logarithmically spaced bands with cut-off frequencies of 200, 362, 655, 1186, 2147, 3885, and 7032 Hz. Pairs of bands were tailored to correspond quite closely with the formant ranges of the talker (B1+B2 ≈ F1; B3+B4 ≈ F2; B5+B6 ≈ F3 and above, denoted F3′). The amplitude envelope (≤30 Hz) of each band was then extracted by half-wave rectification and used to modulate a Gaussian noise source with the same lower and upper cut-off frequencies; increasing the low-pass corner frequency above 30 Hz does not further improve the intelligibility of NV speech (Souza and Rosen 2009). Each band (N1-N6) was scaled to have the same RMS level as that of the corresponding band in the original speech and the bands were summed to create the modulated noise-band speech analogues.

Extracted-formant (EF) stimuli were created from the original sentences using Praat to estimate automatically from the waveform the frequency contours of the first three formants every 1 ms; a 25-ms-long Gaussian window was used. During phonetic segments with frication the third-formant contour often corresponded to the fricative formant rather than to F3. Gross errors in formant-frequency estimates were hand-corrected using a graphics tablet; amplitude contours corresponding to the corrected frequencies were extracted from spectrograms for each sentence. The frequency and amplitude contours were used to generate three-formant analogues of the sentences by means of simple parallel-formant synthesis, using second-order resonators and an excitation pulse modelled on the glottal waveform (Rosenberg 1971). The pitch was monotonous (F0 = 150 Hz), and the 3-dB bandwidths of F1, F2, and F3 were 50, 70, and 90 Hz, respectively.

Figure 2 near here

Reconstructed-formant (RF) stimuli were created from the NV sentences using a simple procedure designed to retrieve the information about formant frequency and amplitude carried by each pair of bands (i.e., N1+N2 for F1, N3+N4 for F2, N5+N6 for F3′). For each pair, the amplitude contour of the reconstructed formant was computed frame-by-frame as the mean amplitude across both bands. The frequency contour was derived from frame-by-frame changes in the relative amplitudes of the two bands within each pair. Figure 2 depicts the reconstruction of the F2 frequency contour from the band pair N3+N4 for an example sentence. The reconstructed contours were used to generate three-formant analogues of the sentences by parallel synthesis, as described above for the EF stimuli. At a particular time, the implied frequency, F, is given by:

$$\log F = \log(g) + kw\log\left(\frac{f_{hi}}{g}\right), \qquad (1)$$

$$w = \frac{a_{hi} - a_{lo}}{a_{hi} + a_{lo}} \quad (-1 \le w \le +1), \qquad (2)$$

where a_{lo} and a_{hi} are the amplitudes of the lower and upper bands, f_{hi} is the upper cut-off frequency of the upper band, k ($0<k\le1$) is a scale factor determining the maximum possible frequency range, and g is the geometric mean frequency of the lower and upper bands. The value of k used here was 0.9; this was to ensure that the frequency range available for formant excursions in the reconstructions was substantial, but not so great as to have allowed unnaturally close approaches between neighbouring formants. Note that low-pass filtering the original sentences at 5 kHz lowers the amplitude of band N6 in the NV stimuli, which tends to lower the frequency, as well as the amplitude, of the reconstructed F3′, particularly during fricative segments. This improves the overall quality of the RF stimuli by reducing the “buzziness” of these segments.

Constant-formant (CF) stimuli differed from their RF counterparts only in that the frequency of each formant was set to be constant at the geometric mean frequency of the whole reconstructed track. For all conditions, the speech analogues were played at a sample rate of 22.05 kHz and 16-bit resolution over Sennheiser HD 480-13II earphones, via a sound card, programmable attenuators (Tucker-Davis Technologies PA5), and a headphone buffer (TDT HB7). Output levels were calibrated using a sound-level meter (Brüel and Kjaer, type 2209) coupled to the earphones by an artificial ear (type 4153). All stimuli were shaped using 10-ms raised-cosine onset and offset ramps and presented diotically at 75 dB SPL.

(c) Procedure

Listeners were tested whilst seated in front of a computer screen and a keyboard in a sound-attenuating booth. There were two phases to the study, training and the main experiment, which together took less than an hour to complete. Stimuli were presented in quasi-random order in both phases of the study. Listeners first completed a training session to familiarise them with synthetic-formant and noise-vocoded speech analogues, in that order. The former were examples of EF stimuli, but differed from those used in the main experiment in that the natural pitch contour was used in the resynthesis; listeners were not exposed to RF or CF stimuli during training. The stimuli for each part of the training were derived from 40 sentences taken from commercially available recordings of the IEEE sentence lists (IEEE, 1969). On each of the 40 trials in each part, participants were able to listen to the stimulus up to a maximum of six times before typing in their transcription of the sentence. After each transcription was entered, feedback to the listener was provided by playing the original recording followed by a repeat of the speech analogue. Davis et al. (2005) found this "degraded-clear-degraded" presentation strategy to be an efficient way of enhancing the perceptual learning of speech analogues. All listeners who obtained scores of ≥60% keywords correct in the second half of each set of training trials were included in the main experiment. As in the training, participants in the main experiment were able to listen to each stimulus up to six times before typing in their transcription, and the time available to respond was not limited. However, this time the listeners did not receive feedback of any kind on their responses.

(d) Data analysis

For each listener, the intelligibility of each sentence was quantified in terms of the overall percentage of phonetic segments identified correctly. Phonetic scores are usually more effective at distinguishing performance between conditions for which there is limited intelligibility, owing to floor effects in keyword scores. Listeners' typed responses were converted automatically into phonetic representations using eSpeak (Duddington 2008) for comparison with stored phonetic representations of the original sentences. Phonetic scores were computed using HResults, part of the HTK software (Young et al. 2006). HResults uses a string alignment algorithm to find an optimal match between two strings.

3. RESULTS AND DISCUSSION

Figure 3 shows the mean percentage of phonetic segments identified correctly across conditions, with inter-subject standard errors. A within-subjects analysis of variance (ANOVA) showed a highly significant effect of condition on intelligibility [$F(3,57)= 278.9$, $p<0.001$, $\eta^2=0.936$]. Paired-samples comparisons (two-tailed) were computed using the restricted least-significant-difference test (Snedecor and Cochran 1967); the scores for each condition differed significantly from those for every other condition ($p<0.001$, in all cases). Scores were very high for the NV speech, given that there were only six spectral bands (cf. Shannon et al., 1995). This may reflect the tailored alignment of each pair of bands in relation to the talker's ranges of formant frequencies. More generally, the intelligibility of NV speech tends to be lower when the inventory of stimuli is derived from multiple talkers (Loizou et al. 1999); this reflects the need for listeners to accommodate acoustic-phonetic variability across talkers (see, e.g., Mullennix et al., 1989). Scores were somewhat lower for the EF speech, probably as a result of two sources of error in recreating phonetically relevant acoustic detail. First, estimation of formant frequencies from fluent speech is a technical challenge and prone to inaccuracy, even when the output of an algorithm for the automatic extraction of formant frequencies is subject to hand correction. Such errors in the formant-frequency parameters fed to the synthesiser would be expected to impair intelligibility (Assmann and Katz, 2005). Second, the use of a minimal model for the formant synthesiser, incorporating only three fixed-bandwidth formants, will have introduced synthesis errors, notably in the reproduction of phonetic segments having significant amounts of high-frequency energy, such as voiceless fricatives. These sources of error do not contribute to the process of creating NV speech.

Figure 3 near here

Scores approached 30% when the three-formant analogues were created using frequency and amplitude contours that were reconstructed from the amplitude-envelope information carried by the three band pairs comprising the NV analogues of the original speech. Hence, RF speech was still nearly half as intelligible as EF speech, even using such a simplistic approach to reconstructing the formant-frequency contours from the NV speech. The frequency resolution of normal-hearing listeners far exceeds that required to retrieve this information from the representation of noise-vocoded speech in the peripheral auditory system (see, e.g., Moore 2003). Performance was halved again for CF speech, for which the reconstructed frequency contours were replaced with constant values set to the geometric mean of each formant track. At least in part, the non-zero performance for the CF speech might be because the reconstructed amplitude contours still convey useful information about vocal tract dynamics. Note, however, that simulations comparing randomly generated text strings with those specifying the stimuli used here suggest that baseline phonetic scores can be in the region of 15% for entirely unintelligible speech; the mean score for the CF stimuli was 12%. Remez and Rubin (1990) explored the relative contributions of variations in the frequency and amplitude contours of formants to the intelligibility of sine-wave speech, created by adding together pure tones that follow the frequency and amplitude contours of the lower formants (Bailey et al 1977; Remez et al. 1981). They concluded that frequency variation is far more important than amplitude variation for maintaining intelligibility; this is also true for across-formant grouping in sine-wave speech (Roberts et al. 2010).

The higher recognition scores observed for the RF relative to the CF condition support the notion that changes in the relative amplitudes of different bands in NV speech convey useful phonetic information about formant frequency variation. Consistent with this view, the effect of quantising the amplitude envelope into a small number of steps (<8) has a much greater impact on the intelligibility of sine-vocoded speech processed through a small (6) rather than a large (16) number of channels, presumably because the reduced information available from across-channel amplitude comparisons makes it more difficult to infer the underlying formant frequencies (Loizou et al 1999). The importance of combining information across a small number of channels to reconstruct signal properties important for intelligibility is also evident in Apoux and Healy's (2009) demonstration that phonemes can be identified from relatively few randomly selected channels, even when noise is present in other channels. Dorman et al. (1997) suggested that the mechanism mediating the high degree of intelligibility achievable with a small number of channels may be the same as that mediating the recognition of speech produced by talkers with a high fundamental frequency.

Recently, more direct evidence that the frequency contours of formants can be inferred from across-channel amplitude comparisons, at least for single formant transitions, has been provided by Fox et al. (2008). Their study explored the role of F3 transitions in distinguishing the place of articulation of initial stops in the syllable pairs [da]-[ga] and [ta]-[ka]. They compared actual F3 transitions with virtual ones, where the percept of a frequency transition was cued by a dynamic change in spectral centre of gravity over 50 ms arising from a smooth but rapid change in the relative amplitude of two noise-excited formants with constant frequency (1907 Hz and 2861 Hz). These frequencies are easily resolvable by the peripheral auditory system, but fall within the much larger bandwidth of about 3.5 critical bands (roughly 5 equivalent rectangular bandwidths; Glasberg and Moore 1990) over which the central auditory system appears to integrate phonetic information (e.g., Delattre et al. 1952; Carlson et al. 1975; Chistovich 1985). Virtual F3 transitions were broadly comparable with actual F3 transitions in supporting the correct identification of initial stops; listeners could also distinguish the direction of the F3 transitions when heard in isolation as rising or falling, whether actual or virtual. Fox et al. (2008) concluded that amplitude and frequency information can be combined in the perception of formant transitions.

To conclude, across-band comparisons of amplitude envelopes in NV speech can provide phonetically important information about the implied frequency contours of the underlying formants for sentence-length utterances. In principle, this dynamic spectral information is easily accessible to most listeners even when the number of channels available is relatively limited.

Acknowledgements: This work was supported by EPSRC (UK). Grant Reference EP/F016484/1 (Roberts & Bailey).

REFERENCES

Apoux, F., Healy, E.F. 2009 On the number of auditory filter outputs needed to understand speech: Further evidence for auditory channel independence. *Hear. Res.* **255**, 99-108.

Assmann, P.F., Katz, W.F. 2005 Synthesis fidelity and time-varying spectral change in vowels. *J Acoust Soc Am.* **117,** 886-895.

Bailey, P.J., Summerfield, Q., Dorman, M. 1977 On the identification of sine-wave analogues of certain speech sounds. *Haskins Lab. Status Rep.* **SR-51/52,** 1-25.

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

Boersma, P., Weenink, D. 2008 Praat: Doing phonetics by computer [software package], version 5.0.18, retrieved 1 April 2008 from http://www.praat.org/

Carlson, R., Fant, G., Granstrom, B. 1975 Two-formant models, pitch and vowel perception. In *Auditory Analysis and Perception of Speech* (eds. G. Fant, M.A.A. Tatham), pp. 55-82. London: Academic Press.

Chatterjee, M., Peredo, F., Nelson, D., Başkent, D. 2010 Recognition of interrupted sentences under conditions of spectral degradation. *J. Acoust. Soc. Am.* **127,** EL37-EL41.

Chen, F., Loizou, P.C. 2010 Contribution of consonant landmarks to speech recognition in simulated acoustic-electric hearing. *Ear Hear.* **31,** 259-267.

Chistovich, L.A. 1985 Central auditory processing of peripheral vowel spectra. *J. Acoust. Soc. Am.* **77,** 789-805.

Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K., 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.

Delattre, P.C., Liberman, A.M., Cooper, F.S., Gerstman, L.J. 1952 An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word* **8,** 195-210.

Dorman, M., Loizou, P., 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. *J. Acoust. Soc. Am.* **102,** 2403-2411.

Duddington, J. 2008 eSpeak 1.36, http://espeak.sourceforge.net/

Dudley, H. 1939 Remaking speech. *J. Acoust. Soc. Am.* **11,** 169-177.

Eisner, F., McGettigan, C., Faulkner, A., Rosen, S., Scott, S.K. 2010 Inferior frontal gyrus activation predicts individual differences in perceptual learning of cochlear-implant simulations. *J. Neurosci.* **30,** 7179-7186.

Fox, R.A., Jacewicz, E., Feth, L.L. 2008 Spectral integration of dynamic cues in the perception of syllable-initial stops. *Phonetica* **65,** 19-44.

Glasberg, B.R., Moore, B.C.J. 1990 Derivation of auditory filter shapes from notched-noise data. *Hear. Res.* **47,** 103-138.

Hopkins, K., Moore, B.C.J. 2010 The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects. *J. Acoust. Soc. Am.* **127,** 1595-1608.

Institute of Electrical and Electronics Engineers (IEEE) 1969 IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust.* **AU-17,** 225-246.

Li, N., & Loizou, P. C. 2009 Factors affecting masking release in cochlear-implant vocoded speech. *J. Acoust. Soc. Am.* **126,** 338-346.

Loebach, J.L., Pisoni, D.B., Svirsky, M.A. 2009 Transfer of auditory perceptual learning with spectrally reduced speech to speech and nonspeech tasks: Implications for cochlear implants. *Ear Hear.* **30,** 662-674.

Loizou, P., Dorman, M., Powell, V. 1998 The recognition of vowels produced by men, women, boys and girls by cochlear implant patients using a six-channel CIS processor. *J. Acoust. Soc. Am.* **103,** 1141-1149.

Loizou, P.C., Dorman, M., Tu, Z. 1999 On the number of channels needed to understand speech. *J. Acoust. Soc. Am.* **106,** 2097-2103.

Mullennix, J., Pisoni, D., Martin, C. 1989 Some effects of talker variability on spoken word recognition. *J. Acoust. Soc. Am.* **85,** 365-378.

Moore, B.C.J. 2003 *An Introduction to the Psychology of Hearing, 5th ed.* London: Academic Press.

Nittrouer, S., Lowenstein, J.H., Packer, R. 2009 Children discover the spectral skeletons in their native language before the amplitude envelopes. *J. Exp. Psychol. Hum. Percept. Perform.* **35,** 1245-1253.

Remez, R.E., Rubin, P.E. 1990 On the perception of speech from time-varying acoustic information: Contributions of amplitude variation. *Percept. Psychophys.* **48,** 313-325.

Remez, R.E., Rubin, P.E., Pisoni, D.B., Carrell, T.D. 1981 Speech perception without traditional speech cues. *Science* **212,** 947-950.

Roberts, B., Summers, R.J., Bailey, P.J. 2010 The perceptual organization of sine-wave speech under competitive conditions. *J. Acoust. Soc. Am.* **128,** 804-817.

Rosen, S. 1992 Temporal information in speech: acoustic, auditory and linguistic aspects. *Phil. Trans. R. Soc. Lond. B* **336,** 367-373.

Rosenberg, A.E. 1971 Effect of glottal pulse shape on the quality of natural vowels. *J. Acoust. Soc. Am.* **49,** 583-590.

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

Snedecor, G.W., Cochran, W.G. 1967 *Statistical Methods, 6th ed.* Ames, Iowa: Iowa University Press.

Souza, P., Rosen, S. 2009 Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech. *J. Acoust. Soc. Am.* **126,** 792-805.

Xu, L., Pfingst, B.E. 2008 Spectral and temporal cues for speech recognition: Implications for auditory prostheses. *Hear. Res.* **242,** 132-140.

Xu, L., Thompson, C.S., Pfingst, B.E. 2005 Relative contributions of spectral and temporal cues for phoneme recognition. *J. Acoust. Soc. Am.* **117,** 3255-3267.

Xu, L., Zheng, Y. 2007 Spectral and temporal cues for phoneme recognition in noise. *J. Acoust. Soc. Am.* **122,** 1758-1764.

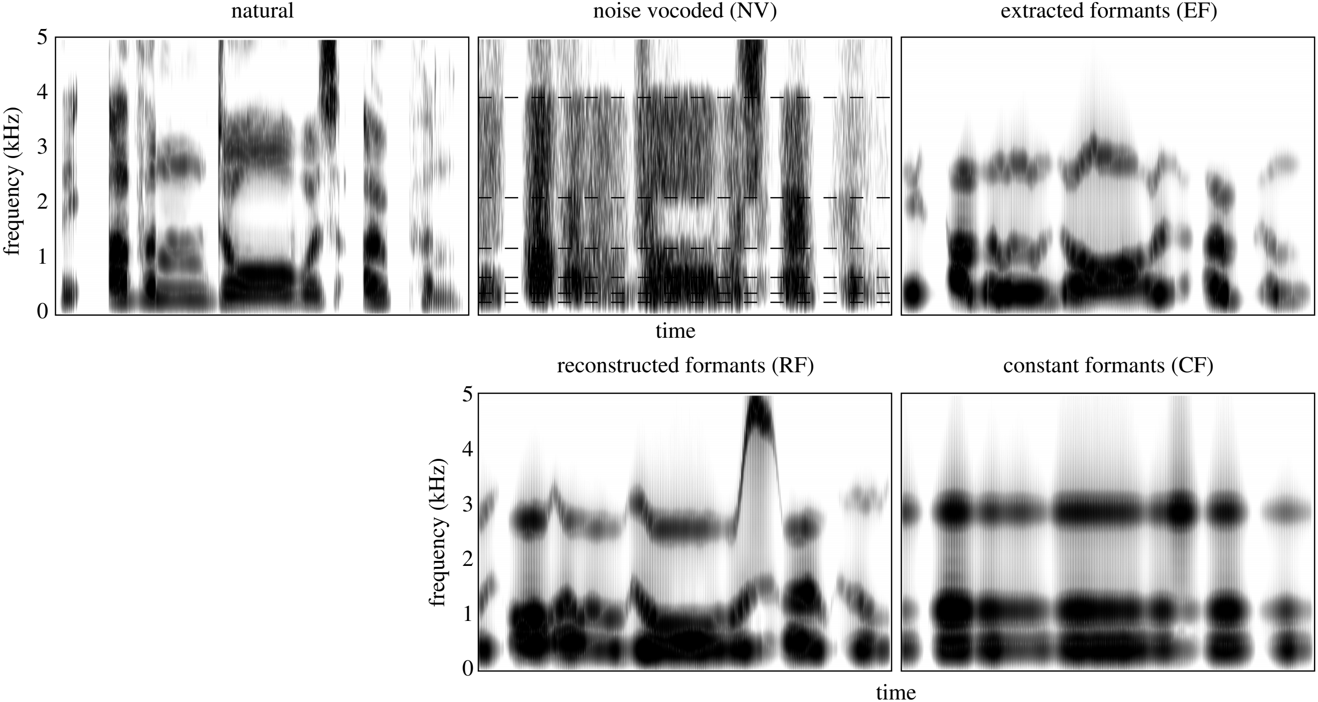
Young, S.J., Evermann, G., Gales, M.J.F., Kershaw, D., Moore, G., Odell, J.J., Ollason, D.G., Povey, D., Valtchev, V., Woodland, P.C. 2006 *The HTK book, version 3.4 Manual.* Cambridge, UK: Department of Engineering, University of Cambridge.

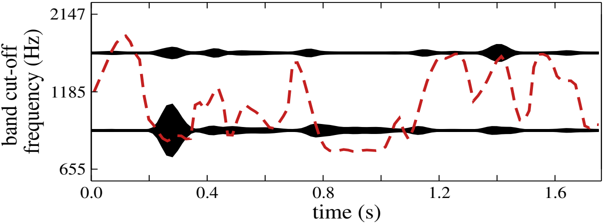
Figure Captions

Figure 1 Spectrograms of an exemplar original sentence, “The oven door was open”, and of the four experimental versions derived from it. The horizontal dashed lines in the panel depicting the noise-vocoded (NV) stimulus indicate the band cut-off frequencies. Note that the most striking discrepancy between the extracted-formants (EF) and the reconstructed-formants (RF) stimuli corresponds to the voiced fricative [z] in “was”. In the EF case, the formant contour extracted by Praat corresponds to F3, but in the RF case the reconstructed formant contour is dominated by the energy in the fricative formant.

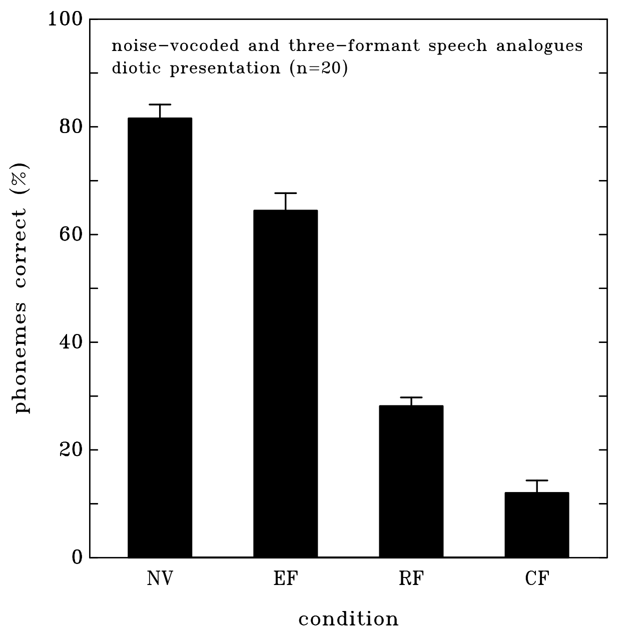
Figure 2 Reconstruction of formant-frequency contours. This schematic illustrates the reconstruction of the frequency contour of F2 from the noise-vocoded (NV) version of the exemplar sentence “The oven door was open”. The reconstructed contour (dashed line) is governed by changes over time in the relative amplitudes of noise bands 3 and 4; the amplitude modulation of each band is depicted by a filled contour centred on the geometric mean frequency. The frequency contour was computed frame by frame using equations 1 and 2 (see main text).

Figure 3 Intelligibility of the four analogues derived from the original sentences. These correspond to the noise-vocoded (NV, 81.6%), extracted-formants (EF, 64.4%), reconstructed-formants (RF, 28.1%), and constant-formants (CF, 12.0%) conditions. Each histogram bar shows the mean phonetic score and corresponding inter-subject standard error (n=20).


natural
noise vocoded (NV)
extracted formants (EF)
reconstructed formants (RF)
constant formants (CF)
frequency (kHz)
0
1
2
3
4
5
time



2147
1185
655
band cut-off frequency (Hz)
0.0
0.4
0.8
1.2
1.6
time (s)


noise-vocoded and three-formant speech analogues
diotic presentation (n=20)
100
80
60
40
20
0
phonemes correct (%)
NV
EF
RF
CF
condition