The relative importance of spectral cues for vowel recognition in severe noise

Rikus Swanepoel, Dirk J. J. Oosthuizen, and Johan J. Hanekom[a)]
Department of Electrical, Electronic and Computer Engineering, University of Pretoria, University Road, Pretoria, 0002, South Africa

(Received 17 February 2011; revised 28 June 2012; accepted 27 August 2012)

The importance of formants and spectral shape was investigated for vowel perception in severe noise. Twelve vowels were synthesized using two different synthesis methods, one where the original spectral detail was preserved, and one where the vowel was represented by the spectral peaks of the first three formants. In addition, formants $F1$ and $F2$ were suppressed individually to investigate the importance of each in severe noise. Vowels were presented to listeners in quiet and in speech-shaped noise at signal to noise ratios (SNRs) of 0, −5, and −10 dB, and vowel confusions were determined in a number of conditions. Results suggest that the auditory system relies on formant information for vowel perception irrespective of the SNR, but that, as noise increases, it relies increasingly on more complete spectral information to perform formant extraction. A second finding was that, while $F2$ is more important in quiet or low noise conditions, $F1$ and $F2$ are of similar importance in severe noise.
[http://dx.doi.org/10.1121/1.4751543]

PACS number(s): 43.71.Es [MAH] Pages: 2652–2662

I. INTRODUCTION

It is known that the introduction of noise to speech affects listening test results in a significant manner (Gong, 1995). Poor speech intelligibility in noise is due to the masking of information that would normally be available to listeners, leaving them with access to only certain parts of the spectral and temporal information in the signal (Juang, 1991; Nabelek *et al.*, 1992). The present study considered vowel recognition in noise, and specifically investigated the importance of formants when listening in severe noise.

Speech recognition becomes extremely difficult at signal to noise ratios (SNRs) lower than 0 dB for noisy environmental conditions like, for example, factory and helicopter noise (Cooke *et al.*, 2001). Previous studies have investigated vowel cues in noise for SNRs of −5 dB and higher (Nabelek *et al.*, 1992; Parikh and Loizou, 2005). The latter authors reported vowel recognition scores as high as 75% (in a set containing 11 vowels) at SNR levels of −5 dB. Many studies have considered the cues that listeners use to recognize vowels in quiet and in noise. Some of these are elaborated on in the paragraphs to follow. However, it appears to be unknown whether these cues are still relevant in severe noise at SNR levels lower than −5 dB.

Two main viewpoints in literature are that the major cues underlying vowel perception (most studies considered quiet conditions) are contained in either the formants (Carlson *et al.*, 1975; Kasturi *et al.*, 2002), or, alternatively, in the spectral shape (Ito *et al.*, 2001), while duration (Hillenbrand *et al.*, 2000), formant movement over time (Neel, 2004), and spectral contrast (Leek *et al.*, 1987) also contribute to vowel recognition.

In the first viewpoint, the first two formants ($F1$ and $F2$) have been regarded as the most important cues to vowel identity in many studies (Sakayori *et al.*, 2002). Note that in this viewpoint, identification of formants from the vowel spectrum may rely on one or more aspects of the information contained in the spectrum. It may be necessary to reliably extract the frequency values or frequency ratios of the lower formants from spectral peaks; alternatively, relative spectral band energies in the formant regions of the spectrum may provide the actual cues that identify the formants. Expectation of the importance of formants in noisy conditions depends on which spectral information is actually required to identify formants. Noise-robust features in speech include the voicing periodicity and temporal modulation structure across frequency bands (Assmann and Summerfield, 2004), while spectral peaks are believed to be less affected by broadband noise than valleys. The latter leads to the expectation that formant information may become more reliable than other spectral cues (e.g., spectral shape) as noise increases if formants are identified from frequency values at the spectral peaks. However, if formants are identified from relative energies in critical formant regions of the spectrum, reliability of formant information may diminish as spectral contrast decreases at higher noise levels. Other cues may be redundant in vowels containing clear formants, while their importance may increase as formant information becomes vague in noise (Iverson and Evans, 2007).

The second, alternative, viewpoint on the cues underlying vowel perception is therefore that listeners do not specifically require formant information, but may use redundant information contained in the complete spectrum to make vowel identity decisions. Some studies support this notion and concluded that listeners may rely more heavily on spectral shape information to recognize vowels (Ito *et al.*, 2001), since it provides a more complete description of a vowel

[a)]Author to whom correspondence should be addressed. Electronic mail: johan.hanekom@up.ac.za

 0001-4966/2012/132(4)/2652/11/$30.00

than formants (Zahorian and Jagharghi, 1993). Others have found no major difference between the two types of cues (formants or whole-spectrum) (Hillenbrand *et al.*, 2006), or have concluded that listeners use both formants and spectral detail in the critical formant regions to perceive vowels (Sakayori *et al.*, 2002).

It is not known whether formants are still important in severe noise, or whether listeners depend more on other cues, where possibly redundancy of cues may provide more robustness against noise. Not many published studies have investigated the relative importance of formants versus spectral shape for vowels in noisy conditions. A first step was taken in the study of Parikh and Loizou (2005) where acoustic analyses were done to investigate the effects of noise on the spectrum of vowels. They concluded that $F1$ and $F2$ were not exclusively used as cues in noise, since $F2$ was severely masked and could not be reliably detected in noise. From the confusion matrices of listening tests in noise, it was concluded that listeners also did not rely solely on spectral shape cues, as vowel pairs not having the same spectral shape were still confused with one another.

The objective of the present study was to investigate the importance of formants as noise increases to severe levels by considering the relative importance of formants and the whole spectrum (or spectral shape) when recognizing vowels in noise. This extends the work of Parikh and Loizou (2005) on the influence of noise on vowel cues to lower SNR levels. As these authors reported, vowel recognition scores may still be high at SNR levels of −5 dB, where they found that listeners still depended partially on formants for vowel recognition. At noise levels lower than −5 dB, the role of formants is, however, still unknown.

The relative importance of formants and spectral shape was investigated in speech-shaped noise under conditions of severe noise. An efficient method to test the importance of formant cues is to use synthesized vowels in listening tests. Vowel synthesis allows control over acoustic parameters, permitting the manipulation of selected cues in the process. A similar experiment has been reported in quiet (Hillenbrand *et al.*, 2006).

In the formants-only representation all spectral detail apart from the formant frequencies was disregarded—a spectrally sparse representation of the vowel. In the particular viewpoint adopted, formant frequency is regarded as the primary determinant of vowel identity. While formant amplitude is regarded as also contributing to vowel identification (Kiefte *et al.*, 2010), formant bandwidth is regarded as being unimportant in this context (Carlson *et al.*, 1979). The spectrally rich whole-spectrum vowel representation retained spectral shape information in both spectral valley and spectral peak regions, so that the relative spectral band energies were retained across the spectrum. In other words, from a slightly different perspective, the latter (whole-spectrum) was a band energy representation of the vowel spectrum, while the former (formants-only) representation focused on preservation of frequency information at specific points in the spectrum (i.e., at the formants).

Generation of vowel stimuli follows the methods of Hillenbrand *et al.* (2006), where two source-filter synthesizers were used to synthesize vowels either using the first three formants, or by using the complete spectral envelope. The synthesized vowels were also spectrally manipulated so that $F1$ and $F2$ were suppressed alternately to investigate the relative importance of these formants in severe noise.

II. METHODS

A. Speech material

Speech material consisted of synthetic vowels derived from existing natural vowel utterance recordings. Twelve recorded vowels were used to serve as reference tokens from which the synthetic vowels were derived as described in the following. These particular vowel recordings have been used to measure vowel intelligibility in previous studies; details on the recordings appear in Pretorius *et al.* (2006). Briefly, vowels were elicited in p-/vowel/-t context. These were /ɑ:/, /ɑ/, /æ/, /ɛ/, /e/, /ɛ:/, /i/, /ə/, /u/, /ɔ/, /œ/, and /y:/. Two native Afrikaans speakers in their twenties, one male and one female, repeated each vowel three times. Speakers were selected for clear articulation, and recordings were validated by measuring vowel recognition in quiet to confirm that no vowels were confused in quiet.

One vowel token was selected from the three recorded tokens of each vowel in the set as follows. Vowel tokens were analyzed in 32 ms time windows using MATLAB. Time windows were weighted by a Hamming window to prevent spectral leakage. A time window in the middle of each vowel token was selected to which a 12th-order linear predictive coding (LPC) analysis was applied to produce an LPC spectrum. A peak-picking algorithm was then applied to identify the first three formants. Correct identification of formants was confirmed by comparing results to LPC-based formant analysis of the same speech tokens with PRAAT (a speech analysis package, www.praat.org) (Boersma, 2001). From the three recordings of each vowel, the token with the highest average formant spectral contrast (formant peak to subsequent valley ratio) was selected as the utterances to be used in the experiments.

B. Vowel synthesis

Vowels were synthesized using either a whole-spectrum filter function or a formants-only filter function driven by a glottal pulse signal. The objective in the former was to retain spectral shape detail in the synthesized spectrum. In the latter, the objective was to place spectral components at the formant positions. In both methods, the driving signal used as input to the synthesis filter was a glottal pulse signal.

1. The source-filter model

Vowels were synthesized using a popular model for speech synthesis consisting of a source signal mimicking the vocal cords that is filtered by a finite impulse response (FIR) filter to produce the synthesized speech sound (Paul, 1981). The source is typically single-sample pulses with its period equal to the instantaneous fundamental frequency of the signal being synthesized (Hillenbrand *et al.*, 2006). Klatt (1987) developed a voicing source somewhat different than this simplified source, described by

$$U_g(t) = at^2 - bt^3, \quad (1)$$

where U_g depicts the open phase of a glottal cycle with a and b depending on the amplitude of voicing and the glottal open period. This source, which was incorporated into the KLSYN88 formant synthesizer (Klatt and Klatt, 1990), was used in this study.

The pitch period of the source was set to the instantaneous fundamental period of the original recorded vowel so that the newly created vowel sound was similar to the original spoken vowel. The fundamental frequency was calculated with the Simplified Inverse Filter Tracking (SIFT) algorithm for each time window (Markel, 1972). The source waveform was divided into 32 ms windows with 16 ms overlap, after which each time window was multiplied by a Hamming window to minimize spectral leakage. Vowels were synthesized by filtering each time window of the source signal with either a whole-spectrum or formants-only FIR filter, described in the following. Note that there is no particular significance in the choice of a FIR filter. Once a smoothed whole-spectrum or formants-only spectral shape has been extracted, the synthesizer recreates this spectral shape by an appropriate FIR filter.

2. Whole-spectrum filter function

To obtain a filter response representing the complete spectrum of a vowel, the method by Zahorian and Jagharghi (1993) was used where the coefficients in a discrete cosine transform (DCT) expansion were calculated to provide a smooth approximation of the vowel spectrum. These coefficients are similar to traditional cepstral coefficients, but are referred to as DCT coefficients since they are calculated by taking the cosine expansion of the magnitude spectrum of the signal being analyzed.

A 512-point fast Fourier transform (FFT) was computed for each Hamming-windowed speech frame $x(n)$ of the original recorded vowel sound, after which the logarithm of the magnitude spectrum was taken to compute $X(n)$ as follows:

$$X(n) = \log\Big(|\mathrm{FFT}(x(n))|\Big). \quad (2)$$

The DCT coefficients were calculated as

$$a_m = \frac{2}{N} k_m \sum_{n=0}^{N-1} X(n) \cos\left[\frac{(2n+1)(m-1)\pi}{2N}\right]$$
$$m = 1, \ldots, N, \quad (3)$$

with

$$k_m = \begin{cases} \frac{1}{\sqrt{2}}, & m = 1 \\ 1, & m \neq 1 \end{cases}. \quad (4)$$

The coefficients are depicted by a_m, while N represents the number of coefficients required (Watson and Harrington, 1999). The smoothness of the spectrum depends on the number of coefficients used in the DCT expansion [(Eq. (5)]. As more coefficients are added to the expansion, more spectral detail is modeled. The final smoothed spectrum was obtained by using the first 12 coefficients (a typical choice in speech recognition applications) in the DCT expansion,

$$[H'(f)] = \sum_{m=1}^{N} a_m \cos[(m-1)\pi f], \quad (5)$$

where $[H'(f)]$ represents the frequency response with logarithmic scaled amplitude over a selected frequency range (0–8000 Hz), and $N = 12$ the order of the DCT (Nossair and Zahorian, 1991). This synthesis method provided a smoothed representation of the original vowel spectrum as shown in the example in Fig. 1, but retained spectral shape information across the spectrum.

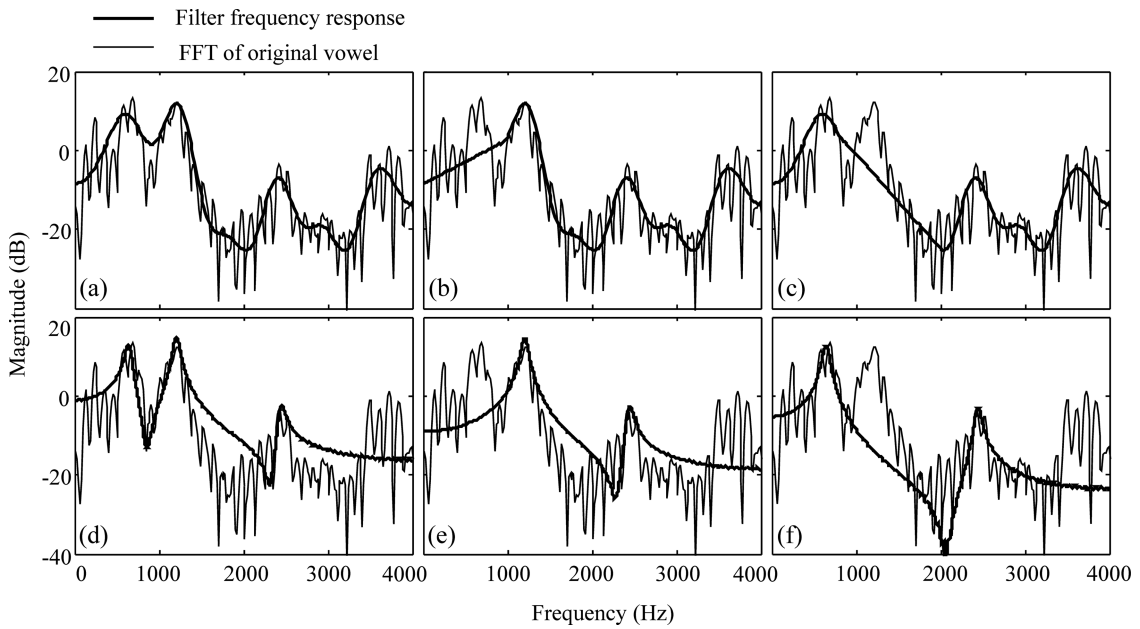

FIG. 1. Filter frequency responses for the whole-spectrum vowels [(a)–(c)] and formants-only vowels [(d)–(f)] of the vowel /ɑ/ (male speaker). The graphs labeled (a) and (d) refer to the response for no formant suppression, while (b) and (e) show the suppression of $F1$ and (c) and (f) the suppression of $F2$.

3. Formants-only filter function

A damped sine wave synthesizer (DSS) (Hillenbrand *et al.*, 2006; Hillenbrand and Houde, 2002) was used to synthesize vowels containing only the first three formants, while other spectral details were suppressed. The DSS creates an impulse response by summing three exponentially damped sinusoids at frequencies, amplitudes, and bandwidths corresponding to the formant values of the natural vowel to be synthesized. The sinusoids are described by

$$d(t) = ae^{-bt\pi}\sin(2\pi ft), \quad t \geq 0, \tag{6}$$

where a is the amplitude, b the bandwidth, and f the frequency of the formant. For the present study, b was held constant at 80 Hz as was done in Hillenbrand *et al.* (2006). For each vowel, formant frequencies and amplitudes were obtained from 12th-order LPC spectra taken over 32 ms Hamming-windowed segments. The final impulse response for each frame was obtained by summing the three damped sinusoids according to the $F1$, $F2$, and $F3$ values. This synthesis method ignored all other spectral detail of the original vowel, including formant bandwidths, and reconstructed vowel spectra with three damped sinusoids. In other words, this is a coarse approximation to the original spectrum with the only details of the resynthesized spectrum corresponding to the original being the formant frequencies and amplitudes of the first three formants. Although Fig. 1 appears to suggest that some spectral shape information is retained, this would be coincidental in any examples where this may occur; in this spectral reconstruction with only three damped sinusoids, no information on the original vowel spectra other than the selected formant amplitudes and frequencies would be available to a listener. Other spectral details that appear have no relation to the original vowel spectrum.

Note that the example in Fig. 1 also appears to suggest that the whole-spectrum representation contains spectral detail of higher frequencies than the formants-only representation. This observation is correct for the example shown, where the whole-spectrum representation contains $F4$ information, but is not true for all vowels. There will be examples of vowels (those with lower $F3$) where the whole-spectrum representation contains information on $F4$, while the formants-only representation will always model the first three formants only. The whole-spectrum representation always retains more spectral detail than the formants-only representation, and where $F4$ is below 4 kHz it will contain $F4$ information. However, perceptual differences (differences in recognition rate in noise) between the whole-spectrum and formants-only representations should not depend on the presence or absence of $F4$; it is well-documented in literature that the lower formant regions, especially $F1$ and $F2$, are most important for vowel recognition (Sakayori *et al.*, 2002). In the context of the present study, the main interest is in whether the additional spectral detail across the entire spectrum in the whole-spectrum representation assists with vowel recognition in noise, or whether the auditory system relies entirely on the formants, while other spectral detail is of less importance.

C. Suppression of F1 and F2

To assess the relative importance of individual formants for vowel recognition in noise, $F1$ and $F2$ were suppressed alternatively for each synthetic vowel stimulus. To suppress these formants, the damped sine wave impulse response of the formants-only vowels was generated either without $F1$ or $F2$ frequency inputs, while the whole-spectrum filter transform function was manually manipulated so that a linear amplitude transition from the start to the end of the suppressed region existed.

Figure 1 shows the filter frequency responses for the whole-spectrum vowels (top row) and formants-only vowels (bottom row) for the vowel /ɑ/ (male speaker). The first column shows the response for the complete spectrum vowels (containing both $F1$ and $F2$), while the second and third columns depict the suppression of $F1$ and $F2$, respectively. It can be seen that the DCT spectral shape follows the original spectrum, while the DSS spectrum matches formant amplitudes and frequencies of the original vowel spectrum, but none of the other spectral detail.

D. Consonants synthesis

As mentioned earlier, formant transitions within the vowel contribute to vowel perception (Neel, 2004). Vowel synthesis was carried out window by window so that duration and formant movement of the original vowel was retained.

The original /p/ and /t/ consonants of the recorded vowels were not integrated with the newly synthesized vowels, as natural consonants contain transitional formant cues that would aid listeners in identifying the vowels (Strange *et al.*, 1983). As vowels were dissected from the utterance and replaced by synthesized formants-only or whole-spectrum representations, it was not clear whether these transitional cues contained in the original consonants would give one type of synthesized vowel an advantage over the other. Also, formant transitions contained in the consonants may negate the effect of suppression of $F1$ or $F2$ in the vowel portion of the synthesized utterance: Vowels may be recognized purely from transitions if the steady-state part of the vowel is absent.

Therefore, while these consonant transitions may enhance vowel identification in noise, this was not part of the investigation in the present study. To ensure that no formant transition cues were transmitted by the consonants, neutral consonants containing no vowel-specific cues were synthesized and concatenated with the synthesized vowel signals. The /p/ consonant was created by using the method described in Hillenbrand *et al.* (2006) to synthesize unvoiced speech. A source signal was generated, consisting of a sequence of pulses with a probability of 0.5 to have amplitude of zero or non-zero at each sampling point. This source signal, which was spectrally indistinguishable from white Gaussian noise, was filtered by the average spectral envelope of the /p/ consonants of the original vowels. The /t/ consonants were created by averaging the time-domain waveforms of all the /t/ consonants, thereby eliminating any vowel cues contained in the consonants.

TABLE I. Concordance index at different noise levels.

SNR (dB)	Concordance index (matrix diagonal included)	Concordance index (confusions only)
−3	1.00	0.48
−6	0.89	0.33
−8	0.82	0.17
−9	0.85	0.12
−10	0.80	0.11
−11	0.77	0.08
−12	0.44	0.12
−13	0.19	0.07

E. Additive noise

Speech-shaped noise was used to ensure an equal SNR at all frequency locations in the spectrum (Phatak and Allen, 2007). The average speech spectrum for vowels containing no $F1$ or $F2$ differs from vowels containing all formants, and therefore three different shapes of noise signals were generated. The first noise signal was generated for the vowels containing both the first and second formants, while the second and third noise signals were generated for the $F1$- and $F2$-suppressed vowels, respectively. A FIR filter, which was derived from the average LPC spectrum related to the specific group of vowels, was used to filter white noise to generate speech-shaped noise.

F. Participants

A group of 12 listeners, 6 males and 6 females, participated in the experimental study. Listeners were native Afrikaans-speaking, had normal hearing determined through screening by an audiologist, and were between the ages of 20 and 28.

G. Noise levels tested

Agreement between listeners in terms of confusion patterns determined the noise levels tested. Pilot vowel confusion data were measured at a number of noise levels (12 listeners, using 12 original vowels to which speech-shaped noise was added, at least 8 repetitions at each noise level). Correspondence between the individual confusion matrices of the listeners was calculated at a number of noise levels with the within-stimulus concordance index (Brusco, 2004). These data are shown in Table I. The concordance index ranges from 0 to 1. Concordance indices are tabulated for two analyses: The first includes the diagonal of the confusion matrices, while the second does not. The latter considers purely the agreement in terms of confusion patterns. A knee appears in the curve of the former at –10 dB SNR, from where agreement between listeners declines rapidly. This knee was used as the cutoff point for the SNRs used in the complete data set.

Noise was added to each synthesized vowel at SNRs of 0, −5, and −10 dB. The final noisy vowels were scaled to an intensity level of 70 dB SPL. Each vowel was centered in an equal duration noise burst of 1000 ms.

H. Procedure

A double-walled sound booth was used as the location for the tests to minimize external noise or interference. A computer, equipped with an M-Audio Fast Track Pro soundcard (M-Audio, Irwindale, CA), was used in the experiments. The speech samples were presented through an M-Audio EX66 loudspeaker (M-Audio, Irwindale, CA) having a frequency response of 37 Hz–22 kHz with ±1 dB passband flatness.

Prior to the main test, each listener completed a practice round where the original whole-spectrum vowels were presented in quiet. The objective of this session was to familiarize listeners with each vowel sound and its description that appeared on the computer screen. Stimuli were then presented in random order to eliminate any predictability. The 12 vowels were presented 10 times (5 male voice and 5 female voice) in 24 conditions, comprising two vowel synthesis types (formants-only and whole-spectrum vowels), three spectral manipulations (complete spectrum, $F1$-suppressed and $F2$-suppressed vowels) and four SNRs (−10, −5, 0 dB, and quiet). In other words, a total number of 2880 utterances were presented to each listener, and each test took approximately 2 h to complete.

III. RESULTS

Mean percentage correct scores and standard deviations across the 12 listeners are documented in Table II for all conditions. A three-way repeated measures analysis of variance (ANOVA) was used to evaluate the effects of synthesis type (ST), spectral manipulation (SM), and SNR. All factors as well as interactions were found to be statistically significant: ST $[F(1,11) = 131.4;\ p < 0.0001]$, SM $[F(2,22) = 242.5;$

TABLE II. Mean and standard deviation values for vowel recognition scores for all conditions.

		Whole-spectrum synthesized vowels				Formants-only synthesized vowels			
		−10 dB	−5 dB	0 dB	Quiet	−10 dB	−5 dB	0 dB	Quiet
Complete spectrum vowel	Mean	62.5	76.5	78.3	84.2	48.9	71.9	84.7	89.3
	s.d.[a]	9.3	5.6	5.5	5	8	9.2	5.4	5.7
$F1$-suppressed	Mean	39.6	56.2	65.3	69.6	31.7	47.3	61.2	64.8
	s.d.	9	8.3	8.5	10.9	7.9	7.6	8.5	7.6
$F2$-suppressed	Mean	34.2	51.6	57.8	65.4	25.3	40.8	50.7	39.9
	s.d.	7.3	7.6	6.7	6.7	4.8	9.6	9.3	7.8

[a]Standard deviation.

$p < 0.0001$], SNR [$F(3,33) = 336$; $p < 0.0001$], ST × SM [$F(2,22) = 63$; $p < 0.0001$], ST × SNR [$F(3,33) = 15.4$; $p < 0.0001$], SM × SNR [$F(6,66) = 6.2$; $p < 0.0001$], ST × SM × SNR [$F(2.9,31.7) = 24$; $p < 0.0001$]. *Post hoc* tests using Bonferroni adjustment showed a significant main effect of the SNR between all combinations of SNRs ($p < 0.0001$) except between 0 dB and quiet ($p > 0.05$). Vowel perception in noise at a SNR of 0 dB therefore remained essentially at the same level as in quiet.

A. Vowel synthesis type

Figure 2 (top panel) depicts the percentage correct scores as bar plots (error bars indicate standard deviation) for the vowels without any spectral manipulation, comparing the scores for the whole-spectrum and formants-only synthesized vowels. The whole-spectrum vowels yielded recognition scores significantly higher than the formants-only vowels at −10 dB SNR [$F(1,22) = 14.83$; $p = 0.001$], while the formants-only vowel scores were higher than the whole-spectrum vowel scores at 0 dB SNR [$F(1,22) = 8.2$; $p < 0.01$] and in quiet [$F(1,22) = 5.52$; $p < 0.05$]. When comparing vowel score differences between the quiet and −10 dB SNR condition, the formants-only vowel scores showed a 39.4% decrease in vowel recognition while the whole-spectrum vowel scores decreased by 21.7%.

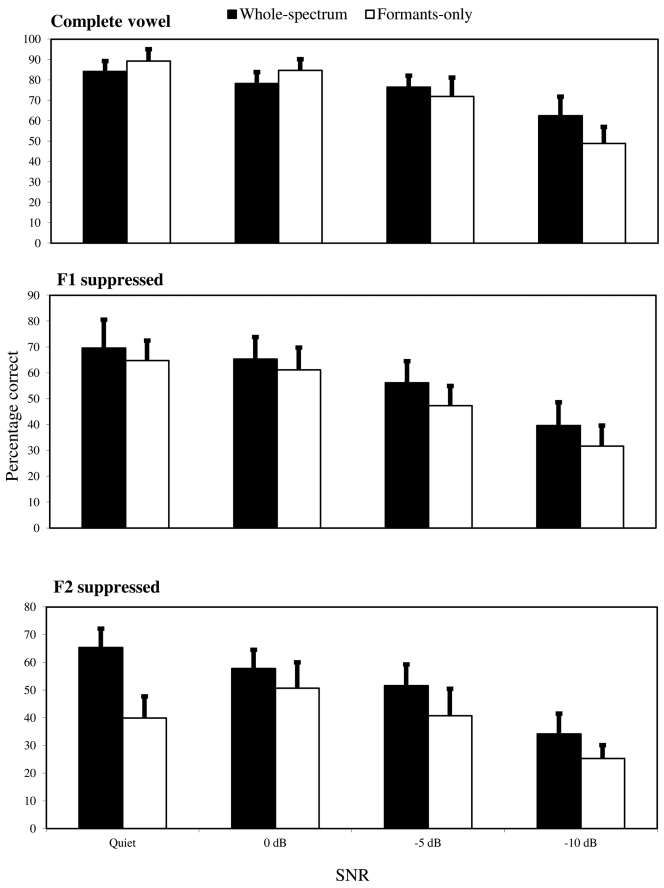

FIG. 2. Listening test results for all listening conditions, comparing whole-spectrum and formants-only vowel scores for stimuli without spectral manipulation (i.e., formants were not suppressed, top panel) as well as stimuli with $F1$ and $F2$ suppressed (bottom two panels).

Apart from these group differences between whole-spectrum and formants-only synthesized vowels, individual differences between vowels were also evident, with particular vowels being more robust against noise. Consideration of individual vowel scores revealed that the better overall recognition of the formants-only vowels over the whole-spectrum vowels at low noise levels (quiet and 0 dB SNR conditions) was primarily due to the scores of the vowels /ɛ:/ and /ɛ/. In severe noise, however, vowels /æ/, /e/, /i/, /u/, /ɛ/, /œ/, and /ɑ/ yielded better recognition for the whole-spectrum vowels than for the formants-only vowels. The two back vowels /ɑ:/ and /ɑ:/ proved to be the most robust in severe noise.

Specific error patterns vary across subjects and agreement between listeners (in terms of confusion patterns) declines as noise increases. Confusion matrices for the vowels in noise (without formant suppression) show that in low noise conditions confusions mostly occur between vowels that are closely spaced in formant space or spectral-band space (defined in the following). In other words, these vowels have similar $F1$, $F2$, duration or spectral band cues, with examples being confusions between /ɛ/ and /i/, and /ə/ and /œ/. Vowels /e/, /ɛ:/, /ɛ/, /ə /, /y:/, and /œ / are confused with vowels situated close together in the spectral-band space. All of these vowels also share a similar $F1$, $F2$ or both formants with the vowels being confused. The trend that vowel confusions are related to Euclidean distance in the two vowel spaces continues at higher noise levels, but random (or apparently random) confusions increase. In severe noise, the longer-duration vowels /ɑ:/, /ɛ:/, /e/, and /y:/ are confused with both shorter and longer duration vowels, while shorter-duration vowels are mainly confused with other short-duration vowels.

B. Suppression of formants

Confusion matrices for the vowels with $F1$ and $F2$ suppressed show that confusions occur between vowels sharing similar cues that are still available after formant suppression. For the $F1$-suppressed vowels in quiet and 0 dB SNR, confusions occur near the diagonal, showing that vowels are confused due to similar $F2$ cues, while near-diagonal confusions for the $F2$-suppressed vowels show that vowels are confused due to similar $F1$ cues. With the addition of noise, confusions along the diagonal increase for both the $F1$-suppressed and $F2$-suppressed vowels. In severe noise, more confusions (other than those along the matrix diagonal) occur for the $F2$-suppressed vowels than for the $F1$-suppressed vowels.

Tukey multiple comparison tests indicated that the scores for the whole-spectrum vowels were significantly higher than the $F1$- and $F2$-suppressed vowels at all SNRs ($p < 0.05$). Vowel scores for the $F1$-suppressed vowels were significantly higher than the $F2$-suppressed vowels in low noise conditions (0 dB SNR and quiet, $p < 0.05$), while in severe noise, the suppression of $F1$ or $F2$ had a similar detrimental effect. The effect of either $F1$ or $F2$ suppression therefore led to a significant reduction in vowel recognition scores in severe noise, while in the quieter conditions, $F2$ was more important in conveying vowel information than $F1$.

For the $F1$-suppressed vowels, the whole-spectrum vowel scores were significantly higher than the formants-only vowel scores at -10 dB [$F(1,22) = 5.25$; $p < 0.05$] and -5 dB [$F(1,22) = 7.48$; $p < 0.05$], while no significant differences were found for the quieter listening conditions. For the $F2$-suppressed vowels, the whole-spectrum vowel scores were significantly higher than the formants-only vowel scores at all SNRs ($p < 0.05$ for 10, -5, 0 dB SNR and quiet) [$F(1,22) = 12.52$, 9.41, 4.6, 73.76]. These results suggest that (1) more robust cues for vowel identity may be embedded in the whole-spectrum representation, as opposed to the sparse formants-only spectrum and (2) the importance of having more complete spectral information of vowels available grows relative to the formants-only representation of vowels as noise increases.

C. Analysis of data with multidimensional scaling

Although the above-presented results show the value of having more complete spectral information available when perceiving vowels in noise, this does not resolve the question of whether the auditory system extracts formant information from the available spectral information (in these experiments, either the whole-spectrum or the formants-only representation), or whether it relies on spectral shape information to recognize vowels.

With the objective of determining whether vowel recognition and confusion patterns can be explained best by a vowel cue space (or, equivalently, vowel space) spanned by formant axes (formant space) or by axes that characterize the complete spectrum (spectral-band space), a multidimensional scaling (MDS) analysis was carried out. MDS is a statistical technique that converts pairwise (dis)similarities between stimuli to a multidimensional spatial configuration of points, where the distances between points are monotonically (and in some variants, linearly) related to the dissimilarities. In the present application of MDS, these points are the vowels used in the experiments. MDS may be used to analyze, based on observed confusion matrices, the underlying cues that influenced a subject's perception.

Spatial configurations of vowels resulting from MDS analyses of confusion matrices were compared to different vowel cue spaces. The objective was not to search for potential cues that would match the dimensions found by the MDS analysis, but rather to compare the importance of formants to that of spectral shape when perceiving vowels in noise. To this end, a set of appropriate vowel cue spaces were constructed. Specifically, the objective was to find the vowel space representation that provided the smallest fitting error to the MDS dimensions. The implicit assumption is that the best fit reflects the cue set that is more likely used by the auditory system when listening in noisy conditions.

1. Construction of vowel cue spaces

To construct appropriate vowel spaces that may be compared to the results of the MDS analysis, vowels needed to be represented by their formants (formant space) or by a richer representation that reflected the spectral shape. The latter may be achieved by a spectral-band representation. In the predominant vowel description, vowel space is spanned by the first two formants and vowel duration. In the present work, duration was either included or excluded in the various analyses as one of the axes spanning the two vowel spaces (formant space and spectral-band space) considered.

These two types of vowel spaces were constructed as follows. Formants and duration were determined as described in Sec. II, while whole-spectrum information was extracted by filtering the spectrum of the synthesized whole-spectrum vowels into five spectral bands. In Klein *et al.* (1970), vowels were bandpass filtered into a number of spectral bands. For each vowel, the output of each bandpass filter was summed energetically and presented to listeners. The outcome revealed that when vowels were represented by five frequency bands spaced between 500 and 4000 Hz and located two-thirds of an octave from each other, an identification score of 94% was still obtained. Similar spectral bands were therefore used in the present study to locate each vowel in a five-dimensional spectral-band space. The dB sound pressure level (SPL) value for each frequency band (with center frequencies at 445, 680, 1120, 1780, and 2800 Hz) was computed by bandpass filtering the vowel spectrum at each defined spectral band region. The five dB SPL values for each vowel were used in this study to define a whole-spectrum vowel space.

Vowels were therefore located in four different vowel cue spaces (summarized in Table III). These were (1) formant space F with $F1$ and $F2$ cues as dimensions; (2) formant space FD that additionally included duration D as third dimension; (3) spectral-band space B containing five spectral band energy cues; and (4) spectral-band space BD that additionally included duration as a sixth dimension. The synthesized formants-only vowels were represented in either of the two formant spaces, while the synthesized whole-spectrum vowels were represented in the spectral-band spaces.

2. MDS analysis of confusion matrices

It was necessary to group confusion matrices of listeners together due to the differences in vowel confusion patterns found between subjects in severe noise. The 12 confusion

TABLE III. Vowel spaces used in the MDS analysis.

Cue set name	Description	Initial dimensionality	% variance contained in three principal dimensions
F	Formant frequencies	2	100
FD	Formant frequencies and duration	3	100
B	Spectral band energies	5	96.5 (male); 95.4 (female)
BD	Spectral band energies and duration	6	93.4 (male); 91.3 (female)

matrices representing results at −10 dB SNR for all vowel groups were pooled into three groups based on the similarity of the confusion data using the concordance index described earlier. The same pooling was then used at other noise levels as well.

After pooling the confusion data into groups, the pooled confusion matrices were converted to dissimilarity matrices (Klein *et al.*, 1970) using the relationship in the following:

$$d_{ij} = \sum_{k=1}^{12} |c_{ik} - c_{jk}|, \tag{7}$$

where d_{ij} represents the entry in column j of row i of the pooled dissimilarity matrix and c_{ij} represents the entry in column j of row i of the pooled confusion matrix (normalized so that the sum of entries in each row is equal to one).

These dissimilarity matrices were converted to three-dimensional perceptual configurations using individual difference scaling (INDSCAL) analysis (Carroll and Chang, 1970). INDSCAL is a MDS algorithm that performs metric analyses and produces unique results, meaning that the resulting spatial configuration need not be rotated to align MDS dimensions to vowel representations in a given cue space. INDSCAL analyses may be performed using NEWMDSX (Coxon *et al.*, 2005).

The INDSCAL algorithm accepted the three pooled confusion matrices as input and returned four three-dimensional configurations. The first configuration is known as the group space and represents the perceptual configuration of all three pools. The other three configurations are known as the private spaces of each pool. Each private space is a scaled version of the group space, where each pool is represented by individual scaling of the group space axes. The group spaces were used for comparison with the four vowel cue spaces.

INDSCAL analysis was performed for the two synthesized vowel conditions (whole-spectrum and formants-only) with the intention of comparing these to the four vowel cue spaces. As mentioned earlier, the objective was not to search for new cues, but rather to compare the importance of formants to that of spectral shape when perceiving vowels in noise.

3. Reducing vowel cue space dimensionality

All four vowel cue spaces were represented in three dimensions using principal component analysis as in Klein *et al.* (1970). The fitting error of the MDS analysis is sensitive to the number of dimensions with an advantage for fewer dimensions. Converting each cue configuration to the same dimensionality allows comparison between cue configurations by measuring how well they fit the perceptual configurations determined by MDS analysis. Three dimensions were selected, because most of the variance present in each cue set is preserved in three dimensions (see Table III) and INDSCAL configurations are still stable for three dimensions. As the number of dimensions are increased, the number of degrees of freedom in the INDSCAL optimization algorithm increases, which may lead to arbitrariness in the placement of some points (Coxon, 1982).

4. Fitting of MDS dimensions to vowel cue spaces

The perceptual configurations (from the MDS analysis) and vowel cue configurations were standardized by moving the centroid to the origin and dividing each dimension by its standard deviation. A Procrustes analysis was used to fit the cue configurations to the perceptual configurations. This technique finds the optimal fit between two multidimensional configurations in a least squares sense by rigid rotation of one of the configurations. The Euclidean distances were calculated between pairs of tokens in the perceptual configuration and the fitted cue configuration. These distances were averaged across tokens as a measure of goodness of fit. As mentioned before, the objective was to find the vowel space representation with the smallest fitting error to the MDS dimensions.

Figure 3 shows the least square error (LSE) of the fit between the MDS dimensions and the two types of synthesized vowels represented in the four different vowel cue spaces. Higher bars correspond to larger fitting errors.

The LSE for the best perceptual data MDS fit to the cue spaces is shown for the male voice and female voice separately [Figs. 3(a) and 3(b)], as well as the average between the two [Fig. 3(c)]. Note that the pattern of confusions is speaker-dependent: The LSE of the best fit of each of the cue spaces to the perceptual data shows clear differences between the male and the female voice. The observed improvement in the fit of the MDS dimensions to the perceptual data at higher noise levels is not unexpected: With too few vowel confusions, the MDS analysis has little data with which to construct a vowel space; a fair number of confusions are required to enable the MDS algorithm to relate confusions with distances between vowel in the chosen vowel space. No other meaning should be ascribed to the apparent improvement of the MDS fitting error to the data at increased noise levels.

Comparing cue spaces that contain duration as a cue (BD and FD) with cues spaces that do not (B and F), it may appear that a slight advantage exists for a formant space that includes a third dimension (duration) in some conditions, but this is non-significant. A two-factor ANOVA considering the averaged data in Fig. 3(c) (factors SNR and vowel space type) indicates significant differences between the different vowel spaces [$F(3,15) = 4.94$, $p < 0.05$] as well as between the different SNRs [$F(3,15) = 4.52$, $p < 0.05$]. *Post hoc* tests show that differences between vowel spaces appear between the formant space and spectral-band space only at −10 dB SNR, with no significant differences between vowel spaces containing duration and those that do not. At the higher SNRs, there is no significant advantage for either the formants-only or the whole-spectrum vowels. Only in severe noise (−10 dB SNR) does a significant advantage for the complete-spectrum vowels become evident.

IV. DISCUSSION

Normal-hearing listeners only struggle with speech recognition in SNRs lower than 0 dB (Assmann and Summerfield, 2004). This is confirmed by results from the present study, where listeners had little difficulty in recognizing vowels in quiet and 0 dB SNR, while a significant decrease in vowel recognition was found at −5 and −10 dB SNR.

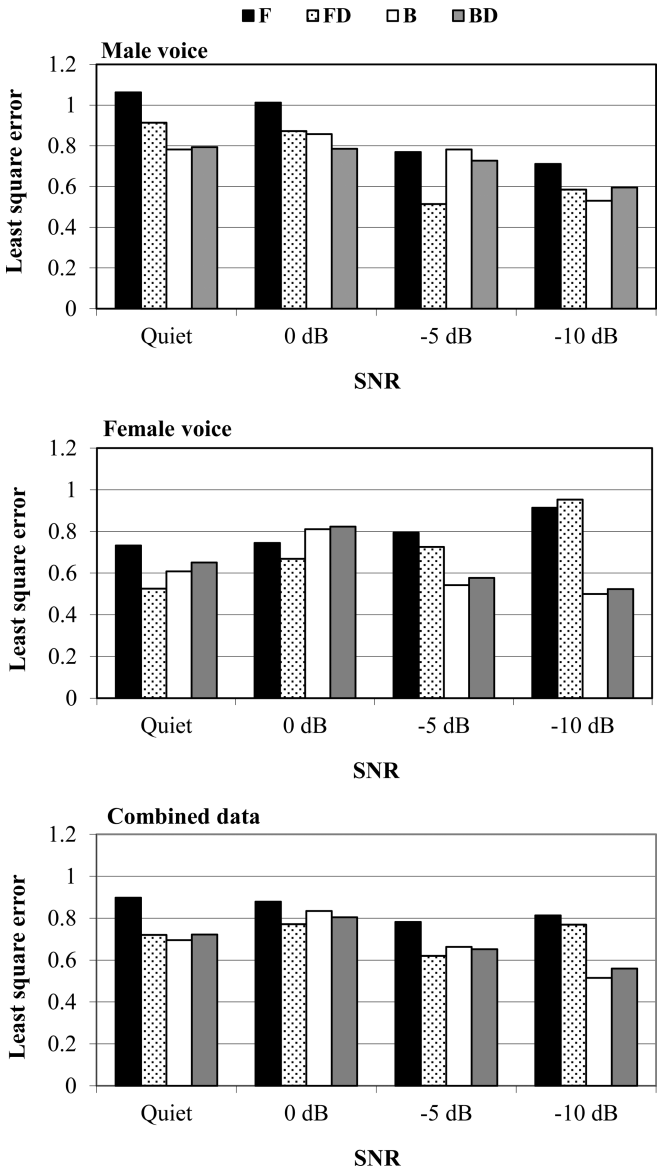

FIG. 3. The least square error fit between MDS dimensions and each of the two vowel synthesis methods (formants-only and whole-spectrum), with vowels represented in either formant space or spectral-band space. As explained in the text, *F* is a formant space with *F*1 and *F*2 as dimensions, while FD has duration as third dimension. B is a spectral-band space with five spectral band energies as dimensions, while BD has duration as sixth dimension.

To reflect on the question whether formants are important for vowel perception in severe noise, bear in mind that the error in spectral synthesis is appreciably lower in the whole-spectrum synthesized vowels than in the formants-only vowels, as is clear from the earlier description of the two synthesis methods. Rather than considering the question "in severe noise, does the auditory system rely on information from the whole spectrum to identify vowels, or from formants only," an equivalent question may be stated as "in noise, would the auditory system perform vowel recognition better with a more accurate spectral representation of the vowels, or is sufficient information available in a more sparse representation (where the error in spectral shape is larger, but where formant information is retained)."

Note first from the MDS analysis that the addition of duration as a dimension does not significantly improve the fitting error of the vowel spaces to the perceptual data, suggesting that duration may not generally be an important cue to vowel identity in severe noise. Of course, if individual vowel confusions were considered, it is possible to reliably discriminate, e.g., /ɑː/ and /ɑ/, but this does not mean that duration is generally a reliable cue for vowel identity. As observed earlier, longer-duration vowels are confused with both shorter- and longer-duration vowels. It appears therefore that noise masks vowel duration cues to some extent; without clear vowel boundaries in noise, auditory mechanisms that estimate vowel duration possibly fare worse for longer vowel sounds so that the value of duration as a cue may decline at higher noise levels.

Figure 2 suggests that the whole-spectrum representation of vowels may become more important as noise increases. Presenting vowels in noise at −5 dB SNR showed no significant difference between vowel identification scores for the whole-spectrum and formants-only vowels, signifying that both types of synthetic vowels contained cues that were sufficient for vowel recognition in noise. As the whole-spectrum vowels also contained formant information, it is conceivable that only the formant peaks were used at this noise level to recognize vowels. For the same noise condition at −5 dB SNR, Parikh and Loizou (2005) concluded that listeners neither use formant cues exclusively, nor distinct spectral shape information to recognize vowels.

At the most severe noise level, −10 dB SNR, vowels containing detailed information on the spectral shape were recognized significantly better than vowels consisting only of formant information. Despite the fact that formant cues alone have been found to be adequate for vowel perception in quiet (Delattre *et al.*, 1952; Klatt, 1982; Molis, 2005), a more complete description of the vowel spectrum in severe noise conditions resulted in better recognition. Zahorian and Jagharghi (1993) proposed that the whole-spectrum representation of a vowel provides some relevant spectral information that is not provided by formants. Sakayori *et al.* (2002) investigated formants and spectral shape combination cues. Even though listening tests were only carried out in quiet, it was found that listeners used the spectral regions at the locations of *F*1 and *F*2 for vowel recognition. Two phonetically different vowels containing similar *F*1 and *F*2 information were found to be easily distinguished, even when the vowels contain spectral information solely at the *F*1 and *F*2 locations. It was argued that the human auditory system identifies formants, but uses spectral shape information in these regions to make a final decision regarding the vowel identity.

A similar argument may be applicable to severe noise conditions. The present data show a significant advantage for the whole-spectrum representation at the lowest SNR, which is also reflected in a better fit of perceptual data to MDS dimensions of a spectral-band space rather than a formant space. However, this was only observed in the most severe noise condition tested; at all other noise levels, information contained in the sparse formants-only representation is sufficient to identify vowels to the same extent as a richer

whole-spectrum representation. This may indicate that the auditory system is able to extract vowel cues from different aspects of the spectrum, i.e., either from the formant peaks (when these are clear) or from spectral band energies. At the most severe noise level tested, formant frequency information may be masked to the extent that formant peaks become unreliable as vowel cues.

However, as noise should be less effective in masking spectral peaks than spectral valleys, this conclusion appears to be counter-intuitive. Returning to the argument in Sec. I, the data may be interpreted as indicating that formants are identified from spectral band energies rather than from frequency positions of spectral peaks. If this were true, it may be expected that noise corruption of the two spectral representations would increase at similar rates as noise levels increase. The data in both Figs. 2 and 3 suggest that this may be true and that a slight advantage may exist for the spectrally richer whole-spectrum representation. The MDS analysis indicates a better fit of a whole-spectrum representation to the perceptual data only at the lowest SNR (significant only for combined male and female voice data at this SNR). This advantage may be because of redundant information in the spectral shape; spectral energies may be more accurately represented in the whole-spectrum representation down to severe noise levels.

Next, consider the relative importance of $F1$ and $F2$. In quiet and in noise at 0 dB SNR, $F2$ was found to be the most salient cue, as the suppression of $F2$ leads to more recognition errors than the suppression of $F1$ for the formants-only vowels. Similar results were obtained in the study by Kasturi *et al.* (2002), where spectral gaps were created for each vowel, and it was established that gaps in the vicinity of $F2$ led to a greater decline in recognition scores than in the $F1$ region. However, results of the present study indicate that these two formants are of similar importance in severe noise.

Perceptual data for vowels in which $F1$ or $F2$ was suppressed appear to support the notion that extraction of spectral band energies underlies identification of the formants. If this were not the case, the expectation would be that without a clear spectral peak at $F1$ or $F2$, the formants-only vowels would lead to significantly weaker vowel recognition performance than the whole-spectrum representation. However, percentage correct scores for whole-spectrum representation and the formants-only representation decline to the same extent when formants are suppressed.

Further support comes from Parikh and Loizou (2005). In listening tests of natural vowels in noise, they found that certain vowels containing similar $F1$ frequencies were not confused with each other, although $F2$ was found to be masked by noise. It was suggested that, in addition to $F1$ information, listeners possibly still obtained information regarding $F2$ by means of spectral shape cues in the $F1$ and $F2$ regions.

In summary, it appears that the auditory system extracts formant information (the MDS fitting error of a formants-only space with perceptual data is similar to that of the whole-spectrum band-energy space), but that it may extract formant information from spectral band energies [possibly in formant regions, as suggested by the work of Sakayori *et al.* (2002)] rather than from formant peaks and corresponding formant frequencies (as the formants-only representation, which should be more noise immune than the whole-spectrum representation if spectral peaks were extracted, was found to fare worse than the latter at high noise levels). At high noise levels, the auditory system may rely more on redundant cues in the whole-spectrum representation to extract formant information.

V. CONCLUSIONS

(1) The MDS fitting error of a formants-only space to the data is generally similar to that of a spectral band's representation at different noise levels, differing significantly only at the most severe noise level tested. This suggests that the task of the auditory system (for vowel perception in quiet and in noise) may be to extract formants from the available spectral information, rather than to identify vowels from the complete spectral shape.
(2) Formant information is possibly extracted from spectral band energies in formant regions rather than from formant frequencies estimated from spectral peak positions.
(3) As noise increases, more formant information is gained from redundancy in the spectral shape in the whole-spectrum representation than from the sparse formants-only representation. Thus, more complete spectral information on vowel identity results in higher vowel recognition levels at high noise levels, while this is not observed at low noise levels.
(4) $F2$ appears to be more important than $F1$ at low noise levels, but these formants appear to be of similar importance at high noise levels.

Assmann, P. F., and Summerfield, Q. (**2004**). "The perception of speech under adverse acoustic conditions," in *Speech Processing in the Auditory system*, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay, Springer Handbook of Auditory Research Vol. 18 (Springer, Berlin).

Boersma, P. (**2001**). "Praat, a system for doing phonetics by computer," Glot. Int. **5**, 341–355.

Brusco, M. J. (**2004**). "On the concordance among empirical confusion matrices for visual and tactual letter recognition," Percept. Psychophys. **66**, 392–397.

Carlson, R., Fant, G., and Granström, B. (**1975**). "Two-formant models, pitch and vowel perception," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tathum (Academic, London), pp. 55–82.

Carlson, R., Granström, B., and Klatt, D. H. (**1979**). "Vowel perception: The relative perceptual salience of selected acoustic manipulations," STL-QPSR **3–4**, 73–83.

Carroll, J., and Chang, J. J. (**1970**). "Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition," Psychometrika **35**, 283–319.

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Commun. **34**, 267–285.

Coxon, A. P. M. (**1982**). *The User's Guide to Multidimensional Scaling* (Heinemann Educational Books, London), p. 59.

Coxon, A. P. M., Brier, A. P., and Hawkins, P. K. (**2005**). "The newMDSX program series," ver. 5, NewMDSX Project, Edinburgh.

Delattre, P., Liberman, A. M., Cooper, F. S., and 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.

Gong, Y. (**1995**). "Speech recognition in noisy environments: A survey," Speech Commun. **16**, 261–291.

Hillenbrand, J. M., Clark, M. J., and Houde, R. A. (**2000**). "Some effects of duration on vowel recognition," J. Acoust. Soc. Am. **108**, 3013–3022.
Hillenbrand, J. M., and Houde, R. A. (**2002**). "Speech synthesis using damped sinusoids," J. Speech Lang. Hear. Res. **45**, 639–650.
Hillenbrand, J. M., Houde, R. A., and Gayvert, R. T. (**2006**). "Speech perception based on spectral peaks versus spectral shape," J. Acoust. Soc. Am. **119**, 4041–4054.
Ito, M., Tsuchida, J., and Yano, M. (**2001**). "On the effectiveness of whole spectral shape for vowel perception," J. Acoust. Soc. Am. **110**, 1141–1149.
Iverson, P., and Evans, B. G. (**2007**). "Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration," J. Acoust. Soc. Am. **122**, 2842–2854.
Juang, B. H. (**1991**). "Speech recognition in adverse environments," Comput. Speech Lang. **5**, 275–294.
Kasturi, K., Loizou, P. C., Dorman, M., and Spahr, T. (**2002**). "The intelligibility of speech with 'holes' in the spectrum," J. Acoust. Soc. Am. **112**, 1102–1111.
Kiefte, M., Enright, T., and Marshall, L. (**2010**). "The role of formant amplitude in the perception of /i/ and /u/," J. Acoust. Soc. Am. **127**, 2611–2621.
Klatt, D. H. (**1982**) "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 1278–1281.
Klatt, D. H. (**1987**). "Review of text-to-speech conversion for English," J. Acoust. Soc. Am. **82**, 737–793.
Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.
Klein, W., Plomp, R., and Pols, L. C. W. (**1970**). "Vowel spectra, vowel spaces and vowel identification," J. Acoust. Soc. Am. **48**, 995–1009.
Leek, M. R., Dorman, M. F., and Summerfield, Q. (**1987**). "Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **81**, 148–154.
Markel, J. D. (**1972**). "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. Audio Electroacoust. **AU-20**, 367–377.
Molis, M. R. (**2005**). "Evaluating models of vowel perception," J. Acoust. Soc. Am. **118**, 1062–1071.
Nabelek, A. K., Czyzewski, Z., and Krishnan, L. A. (**1992**). "The influence of talker differences on vowel identification by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **92**, 1228–1246.
Neel, A. T. (**2004**). "Formant detail needed for vowel identification," ARLO **5**, 125–131.
Nossair, Z. B., and Zahorian, S. A. (**1991**). "Dynamic spectral shape features as acoustic correlates for initial stop consonants," J. Acoust. Soc. Am. **89**, 2978–2991.
Parikh, G., and Loizou, P. C. (**2005**). "The influence of noise on vowel and consonant cues," J. Acoust. Soc. Am. **118**, 3874–3888.
Paul, D. B. (**1981**). "Spectral envelope estimator vocoder," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-29**, 786–794.
Phatak, S. A., and Allen, J. B. (**2007**). "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**, 2312–2326.
Pretorius, L. L., Hanekom, J. J., Van Wieringen, A., and Wouters, J. (**2006**). "n Analitiese tegniek om die foneem-herkenningsvermoë van Suid-Afrikaanse kogleêre inplantingsgebruikers te bepaal" ("Analytical technique to determine the phoneme-recognition ability of South African cochlear implant users"), Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie/S. Afr. J. Sci. Technol. **25**, 195–207.
Sakayori, S., Kitama, T., Chimoto, S., Qin, L., and Sato, Y. (**2002**). "Critical spectral regions for vowel identification," Neurosci. Res. **43**, 155–162.
Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). "Dynamic specification of coarticulated vowels," J. Acoust. Soc. Am. **74**, 695–705.
Watson, C. I., and Harrington, J. (**1999**). "Acoustic evidence for dynamic formant trajectories in Australian English vowels," J. Acoust. Soc. Am. **106**, 458–468.
Zahorian, S. A., and Jagharghi, A. J. (**1993**). "Spectral-shape features versus formants as acoustic correlates for vowels," J. Acoust. Soc. Am. **94**, 1966–1982.