The stimulus duration required to identify vowels, their octave, and their pitch chroma

Ken Robinson[a)]
MRC Institute of Hearing Research, Royal Infirmary, 16 Alexandra Parade, Glasgow G31 2ER, United Kingdom

Roy D. Patterson
MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF, United Kingdom

(Received 7 February 1994; revised 9 May 1995; accepted 10 May 1995)

Computational models of sound segregation typically include the assumption that pitch plays a key role in timbre identification. This hypothesis was investigated by presenting listeners with short segments of static vowel sounds and asking them to identify the vowel quality (timbre), the octave (tone height), or the note (tone chroma) of the sound. There were four vowel categories (/a/, /i/, /u/, and /ɜ/), four octave categories (centered on C1, C2, C3, and C4) and four note categories (C, D, E, and F), and performance was measured as a function of the number of glottal periods of the vowel sound. The results show that at all stimulus durations, it was easiest to identify the vowel quality (mean 94% correct), followed by the octave (71%), and finally the note (52%). The results indicate that timbre can be extracted reliably from segments of vowels that are too short to support equivalent pitch judgments, be they note identification, or the less precise judgment of the octave of the sound. Thus it is unlikely that pitch plays a key role in timbre extraction, at least at short durations. © *1995 Acoustical Society of America.*

PACS numbers: 43.75.Cd, 43.66.Hg

INTRODUCTION

There are theories of auditory perception which suggest that pitch is in some way necessary for the extraction of timbre information from a sound, or that it greatly assists the extraction of timbre information (e.g., Patterson, 1987, p. 178; Terhardt, 1987, p. 279). Informal listening to short segments of vowels and musical notes suggests that one can tell the source of such sounds (their timbre) at durations where the octave and chroma of the sound are not readily apparent, indicating that pitch may not be necessary for timbre processing. In this paper, we review the evidence for the hypothesis that pitch is required for timbre perception, and then present an experiment that measures the stimulus duration required to identify the timbre, the octave, or the pitch chroma of vowel sounds.

A. Pitch assisted timbre identification

A century ago, Stumpf (1890) observed that when two instruments play the same note, their timbres fuse and it is difficult to hear the instruments separately. Later, Cherry (1953) suggested that momentary pitch differences might be one of the cues used to separate simultaneous vowels, and thus voices, at a cocktail party. These observations led Scheffers (1983) to perform concurrent vowel experiments and to construct a computational model to evaluate the hypothesis that the auditory system uses pitch information to direct source segregation and identification. Scheffers (1983) generated pairs of concurrent vowels with the same or different pitches and demonstrated that listeners' performance on a vowel identification task was better when the vowels had differences in fundamental frequency ($f0$) of a semitone or more. He implemented a sophisticated version of the harmonic sieve model of pitch extraction (Goldstein, 1973; Duifhuis *et al.*, 1982), and a vowel identification system based on spectral template matching. Then he attempted to show that vowel identification improved if pairs of pitches were derived from the concurrent vowel on a frame-by-frame basis and used to scale the vowel templates prior to matching. The pitch extraction was typically successful in identifying the $f0$ of one of the concurrent vowels, and in these cases, the template matching was typically successful in identifying the corresponding vowel of the concurrent pair. The pitch extractor was not, however, able to identify the pitch of the second vowel reliably, and when it failed, identification of the second vowel was not reliable either.

Assmann and Summerfield (1990) and Meddis and Hewitt (1992) improved the $f0$ identification for the second vowel by switching from spectral to spectro/temporal pitch extractors. Both groups implemented versions of Licklider's (1951) multichannel, autocorrelation model of pitch, and produced a sequence of autocorrelogram and summary autocorrelograms for each pair of concurrent vowels. An $f0$ estimate was derived from a summary autocorrelogram and used to restrict the information passed to the recognition system to that associated with one $f0$ or the other. The use of autocorrelograms led to better $f0$ extraction for the second vowel, and the segregation of the autocorrelogram information by $f0$ led to better identification of the second vowel. Further improvements to this approach have recently been made by de Cheveigne (1993) and Brown and Cooke (1994). The former extended the harmonic sieve concept to include

[a)]The experiments were performed while author K. R. was a doctoral student at the MRC Applied Psychology Unit (Robinson, 1993).

 0001-4966/95/98(4)/1858/8/$6.00

harmonic cancellation of the components of an interfering source. The latter added harmonic tracking over frames to circumvent the earlier assumption that there were two and only two pitches in the sound.

The computational models in these papers are presented as auditory models, and in each case, they include the assumption that the auditory system can extract an accurate estimate of the absolute value of the $f0$ of a vowel from a single frame of the internal representation of the sound, be it a spectral frame or an autocorrelogram frame. The auditory system is extremely sensitive to changes in the pitch of both sinusoidal and complex tones, provided they are 10 dB above masked threshold (Henning, 1967; Scheffers, 1983). But this does not mean that the auditory system derives an accurate, absolute pitch value from sound segments as short as those represented by the frames in these segregation models (approximately 20 ms in the more recent models).

B. Identification of timbre without the aid of pitch

The alternative position, that pitch is not necessary for the extraction of timbre information, is illustrated by virtually all computational models for machine recognition of speech. The approach dates from at least as early as Klatt (1979), who argued that vowels and, indeed, whole words may be perceived by comparing a sequence of 10-ms spectra with a spectral template. He saw pitch as being useful for later processing, but not of primary importance in the extraction of phonetic information. Over the last decade, the template matching has been replaced by hidden Markov modeling (e.g., Patterson and Hirahara, 1989), but the approach is essentially the same in the sense that $f0$ information is not used to assist phoneme identification which is done from the spectrogram directly.

C. The duration required for timbre, octave, and pitch-chroma identification

1. Timbre

There is some experimental evidence from speech research to support our informal observation that timbre identification is possible even with very short segments of a complex sound. Gray (1942) studied the identification of vowels with $f0$'s ranging from 80 to 384 Hz and durations ranging from 3 to 520 ms. He found that vowels presented as single glottal pulses with durations as short as 5 ms were identified at better than chance levels. Figure 1 shows his data as a function of the number of glottal periods. Vowel identification reaches asymptotic levels in one glottal period for fundamental frequencies from 80 to 192 Hz. Suen and Beddoes (1972) have subsequently confirmed Gray's (1942) findings with digital editing techniques.

2. Octave

Ritsma *et al.* (1965) provide some indirect evidence for the identification of the octave of a sound with short-duration complex sounds. Two listeners were required to match the pitch of a bandpass-filtered pulse train using another pulse train which was bandpass filtered at a different center frequency and presented with a different duration. The fundamental of the pulse train ranged from 70 to 560 Hz in octaves, so there was only one pitch chroma in the experiment and the task was more akin to octave matching. The standard deviation of the pitch matches decreased as the number of cycles increased from 3% of $f0$ at one cycle to 0.5% of $f0$ at seven cycles, indicating that octave information can be extracted from short-duration complex sounds. In a separate experiment where the fundamentals were not in an octave relationship (100 and 333 Hz), the standard deviation at one cycle was well over 10% of the $f0$.

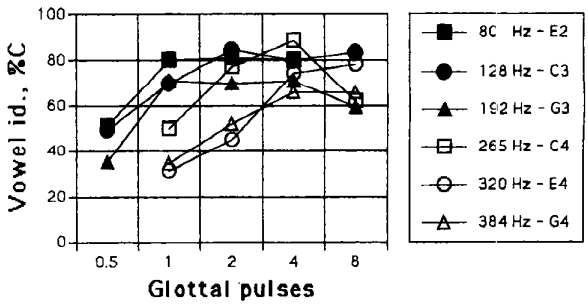

FIG. 1. Vowel identification as a function of the number of glottal pulses for six fundamental frequencies. Data reanalyzed from Gray (1942).

Patterson (1990) and Patterson *et al.* (1993) presented data to show that at long durations, listeners can identify the octave of a note accurately. They used multiharmonic synthetic notes in which the odd harmonics were attenuated from 0 to 27 dB, or the odd harmonics were phase shifted by 0 to 90 deg. The octave of the stimuli ranged from C1 to C6 and the stimuli were presented in a random order. Listeners had to identify the octave of the note either on an integer scale from 1 to 6 (Patterson, 1990), or on a decimal scale from 1.0 to 6.0 (Patterson *et al.*, 1993). Both attenuating and phase shifting the odd harmonics increased the octave rating. For present purposes, however, the primary point is that the average rating for stimuli with a common $f0$ was very close to the physical octave for those stimuli, indicating that for long-duration stimuli listeners can perform an octave identification task reliably.

3. Pitch chroma

Patterson *et al.* (1983) have reported a melodic pitch experiment which does not require pitch identification but which nevertheless seems relevant. A random four-note melody was presented and then repeated, and one of the four notes in the second version was transposed up or down one note of the diatonic scale. The number of cycles was varied to determine when the listeners could identify the position of the transposed note in the second version of the melody. For sinusoids ranging in frequency from 100 to 900 Hz, listeners required about ten cycles to perform the task. The task in this experiment is pitch-interval confirmation which is arguably easier than pitch identification, which suggests that pitch identification might be expected to require stimulus durations longer than those required for timbre identification.

D. Direct comparison of timbre, octave, and pitch-chroma identification

This paper presents a new experimental test of Scheffers' (1983) hypothesis that the auditory system extracts the pitch of a sound on a moment-to-moment basis and uses the pitch value to assist timbre identification. Specifically, we measured the stimulus duration required to identify a vowel, its octave, or its pitch chroma. There were four vowels (/a/, /i/, /ɜ/, and /u/), each of which was synthesized on four notes of the diatonic scale (C, D, E, and F), and all 16 of these stimuli were generated in the octave starting on "middle C" of the keyboard, and in each of the three octaves below middle C, for a total of 64 stimulus conditions. Listeners were required to identify the vowel, the octave, or the note of each stimulus as the number of cycles was varied from 1 to 64. If pitch were essential for timbre perception, we might expect that as we increase stimulus duration, performance on the note identification task would rise above chance either before, or at the same point, as performance for vowel identification.

A note identification task (C, D, E, or F) was chosen in preference to a pitch discrimination task for two reasons: First, it is more comparable to the vowel identification task. Second, the computational models that use pitch to improve vowel identification require an absolute $f0$ value. This is more analogous to a pitch identification task than to a pitch discrimination task. If the computational models of pitch are correct, identifying which of four chroma categories a note belongs to should be a relatively easy task. The auditory models require absolute $f0$ estimates with 2%–3% accuracy. The chroma steps between C, D, and E of the diatonic scale are 12% of $f0$; that from E to F is 6%. The just-noticeable difference for the $f0$ of a musical note would typically be less than 0.5% of $f0$, and so the width of the chroma categories should not restrict performance. The octave identification task was included as an alternative, rather easier, pitch-categorization task, that we knew the listeners would be able to perform from the work of Patterson *et al.* (1993).

I. METHOD

A. Stimuli and equipment

For purposes of this experiment, timbre was specified simply as the quality of the vowels /a/, /i/, /ɜ/, and /u/. The tone height, or octave, of the sound was specified as 1–4 in standard keyboard notation, where middle C is C4, and A4 is 440 Hz. The fundamentals of the notes C1–C4, are just under 33, 66, 131, and 262 Hz, respectively. The tone chroma, or note value, was C, D, E, or F on the equal-temperament scale; that is, each D is 11.9% above the corresponding C, each E is 11.9% above the corresponding D, and each F is 5.9% above the corresponding E. There were a total of 64 conditions in the experiment, and for each, stimuli were generated with 1 to 64 glottal periods.

The vowels /a/, /i/, /ɜ/, and /u/ were digitally synthesized using Klatt's (1980) vowel synthesis program at a sample rate of 8192 Hz. Each vowel was synthesized for four pitch chromas in each of four octaves, giving a total of 16 fundamental frequencies which spanned the range 33–350 Hz. The center frequencies and bandwidths of the four formants in each vowel are detailed in Table I. The formant frequencies and bandwidths were the same as those used by Assman and Summerfield (1989). The vowel stimuli for the experiment were made by excising the fifth glottal period of each Klatt vowel and playing it cyclically 1, 2, 4, 8, 16, 32, or 64 times. The beginning and end of the cycle were both at positive-going zero crossings.

TABLE I. Formant center frequencies and bandwidths for the vowel stimuli used in the experiment.

	Formant Freq. (Hz)	Formant B/W (Hz)
Vowel AH /a/		
formant1	650	60
formant2	950	90
formant3	2950	150
formant4	3300	200
Vowel EE /i/		
formant1	250	60
formant2	2250	90
formant3	3050	150
formant4	3300	200
Vowel ER /ɜ/		
formant1	450	60
formant2	1250	90
formant3	2650	150
formant4	3300	200
Vowel OO /u/		
formant1	250	60
formant2	850	90
formant3	1950	150
formant4	3300	200

The absolute duration of the stimuli varied from 2.9 ms, which is 1 cycle of the F in the highest octave, to 1952 ms, which is 64 cycles of the C in the lowest octave. Within a run, the range of durations was restricted to a factor of 8; that is, the largest number of cycles was 8× the smallest. The variation in duration contributes to the perceptual variability of the set in any given run, and this probably increases the difficulty of the task. Nevertheless, as the perceptual load is the same for all three tasks, the procedure does not make the pitch-chroma task inherently more difficult than the timbre task or the octave task.

The spectrum of each vowel on the note C1 ($f0$ 66 Hz) is presented in Fig. 2. The ordinate shows relative level in dB; the abscissa is frequency. The cutoff frequency of the antialiasing filters was set at 3.3 kHz, and so the higher harmonics shown in Fig. 2 were not presented to the listeners. The upper harmonics are not essential for the identification of vowels, which are principally recognized by the energy in the first two formants (Fant, 1962). The highest formant frequency ($F4$) for the vowel stimuli was 3.3 kHz, and as this was shared by all four vowels, it is unlikely that intelligibility was affected.

Following intensity equalization, the /u/ stimuli were clearly less loud than the other vowels, and the /i/ stimuli were clearly louder. To reduce these and other loudness differences, the /u/ stimuli were increased in level by 3 dB, and

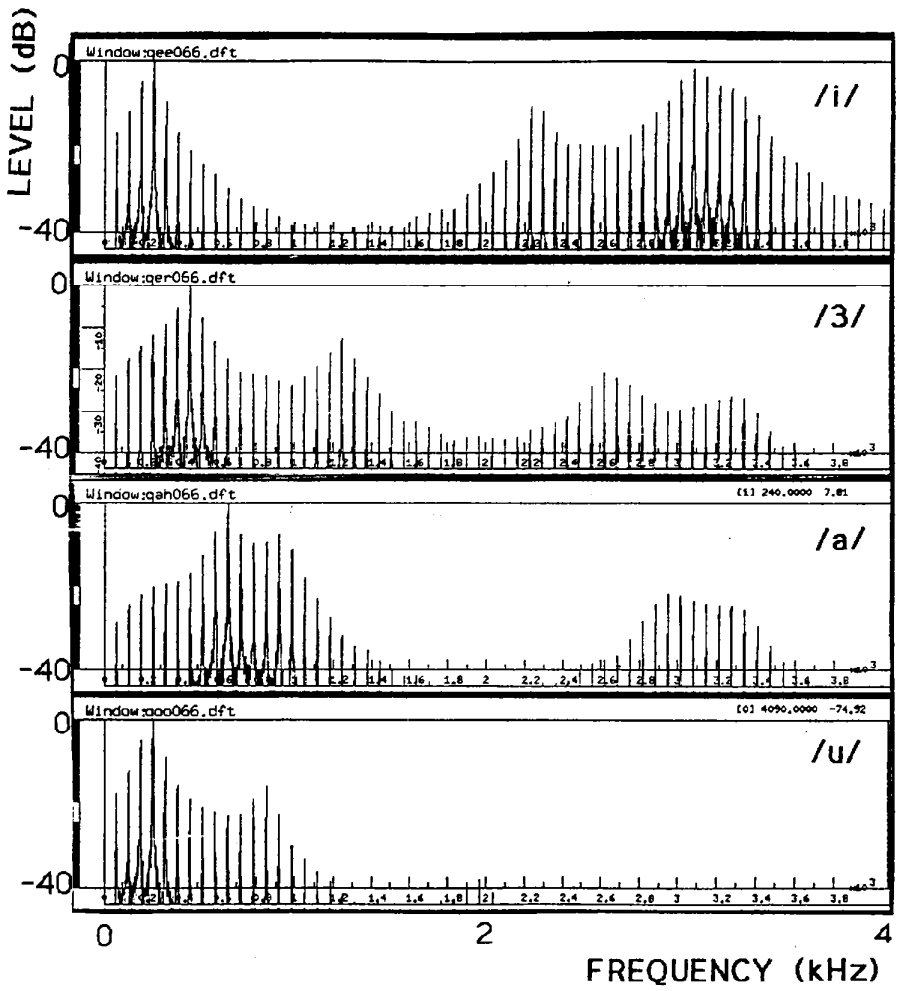

FIG. 2. Amplitude spectra for the vowel stimuli /a/, /i/, /ɜ/, and /u/. The fundamental frequency is 66 Hz (C2).

the /i/ stimuli were decreased in level by 3 dB, all relative to the /a/ stimuli. For each doubling in the number of cycles, the stimuli were reduced in power by 3 dB. The presentation level for the 64-cycle /a/ vowel was 65 dB SPL total power, when played continuously.

The stimuli were presented via a 12-bit digital-to-analog converter, two low-pass filters connected in series (96-dB/oct attenuation above the cutoff frequency), a programmable attenuator, and a Quad 303 amplifier to a pair of Sennheiser HD540R headphones. The fidelity of the system was measured at the input to the headphones by presenting a 1-kHz sinusoid at the level of the vowels with the greatest peak amplitude; specifically, the level was 86 dB SPL for the one-cycle /u/ vowel. The harmonics of the sinusoid and the noise floor were both more than 60 dB down from the level of the sinusoid. The stimuli were presented binaurally to the listener in an IAC sound-attenuated booth.

B. Procedure and listeners

On each trial, the listener was presented a single sound representing one of the 64 combinations of four vowels, four octaves, and four pitch chromas. A trial consisted of a 200-ms ready light, followed by a 300-ms silence, and then a single presentation of the stimulus. Listeners had 5 s in which to respond, and they were given feedback on every trial. There were seven durations (1, 2, 4, 8, 16, 32, and 64 cycles) and in each run of the experiment, all of the stimuli associated with four adjacent durations were presented, for a total of 256 trials. The shortest of these four durations was varied between runs to measure the psychometric function and devote most trials to the steepest part of the psychometric function. Stimulus presentation was randomized in the dimension of interest, and either blocked or randomized across the other two dimensions. For example, when the task was vowel identification in the blocked condition, both vowels and number of cycles were randomly varied between trials, whereas octave and note were blocked. In the randomized condition, vowel, octave, note, and number of cycles were randomized for every trial. The randomization was performed without replacement.

On any given day, the type of response required of the listener was fixed, and there were four response alternatives: For vowel identification, the response buttons were labeled "ah," "ee," "er," or "oo." For octave identification, the re-

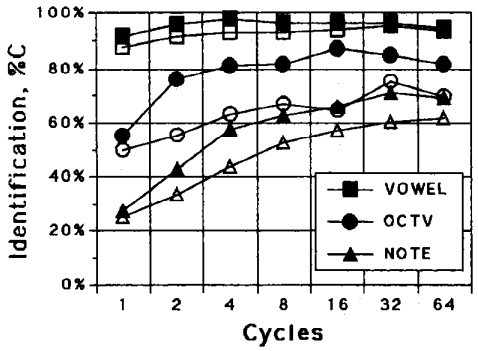

FIG. 3. Group psychometric functions for vowel (square symbols), octave (circular symbols), and note identification (triangular symbols). Filled and unfilled symbols denote blocked and random stimulus presentation.

sponse choices were octaves "1," "2," "3," or "4," and for note identification the available responses were "C," "D," "E," or "F." Listeners were asked to identify vowels on day 1, octaves on day 2, and notes on day 3, and then the task order was reversed for days 4–6. All listeners completed days 1–6; two of the listeners were available for further testing and for them the experiment was replicated over the course of another 6 days. Each day began with a demonstration of the stimuli to remind listeners of the full range. They were then given 32 practice trials on the response task of that day in which no data were collected. During each run of the experiment, 4 demonstration trials were presented with the correct response between every set of 16 trials.

Four listeners participated in each version of the experiment; two participated in both the blocked and random versions. The listeners ranged in age from 18 to 42 yr and all had normal binaural hearing thresholds as tested by pure-tone audiometry at frequencies of 0.25, 0.5, 1, 2, and 4 kHz.

In musical terms, the entire experiment was performed within the key of C major. This means that listeners who understood the concept of the tonic, either explicitly or implicitly, did not need to extract the absolute value of $f0$ to perform the task. They could use an interval judgment to perform the four-way categorization. In psychoacoustic terms, these listeners would be making a four-way discrimination.

In the current experiment, the stimuli are described in terms of the number of cycles of the sound, rather than the duration in milliseconds. The issue as to which measure is more appropriate was investigated in Robinson (1993) who used pulse-train stimuli, a seven octave range of $f0$'s, and a note identification task. At the highest octaves (above 1048 Hz, C6), there was evidence that as $f0$ increased, an increase in the number of cycles was required to maintain the same level of note identification. But in the lowest four octaves (C1–C4), a fixed number of cycles led to a more constant level of performance.

II. RESULTS

The group psychometric functions for vowel, octave, and note identification are presented in Fig. 3. Vowel identification (squares) is about 90% correct with one cycle of the stimulus, and rises to ceiling performance at two cycles. Performance in the randomized conditions is almost as good as that in the blocked conditions. Octave identification is above chance but below ceiling performance with one cycle of the stimulus (about 53% correct). Performance in the blocked conditions rises rapidly to over 80% correct as the number of cycles increases to 64; performance in the randomized conditions rises slowly to about 70% correct. Note that identification is at chance performance with one cycle of the stimulus, and rises steadily as the number of cycles increases to about 70% correct in the blocked conditions and 60% correct in the randomized conditions. An analysis of variance revealed an overall task effect ($F_{2,12}=51.23$, $p<0.001$); vowel identification was easier than octave and note identification (mean performance 94%, 71%, and 52%). *Post hoc* comparisons confirmed that vowels were better identified than octaves and notes, and that octaves were better identified than notes ($p<0.005$). Both octave and note identification gradually improved as the number of cycles increased; vowel identification was essentially independent of the number of cycles. This was reflected in the task by cycle interaction in the analysis of variance ($F_{12,72}=17.92$, $p<0.001$).

The results show that timbre information can be extracted from short stimulus segments that are not sufficiently long to support good pitch identification. They confirm Gray's (1942) finding that vowel identification is possible with a single cycle of the sound. They also confirm the results of Ritsma *et al.* (1965) and Patterson *et al.* (1983); a minimal pitch sensation arises with as little as two cycles, but six to eight cycles are required for stable pitch matching. Performance for octave identification is intermediate between vowel and note identification presumably because it requires a less precise estimate of pitch.

A. Vowel identification

The data for vowel identification were analyzed separately to assess the interaction of vowel quality with octave and note value. There was an effect of increasing $f0$ which appeared as a significant interaction with octave and note ($F_{9,54}=35.57$, $p<0.001$). The interaction was particularly evident for the notes in octave 4; as the note increased from C4 to F4, performance dropped from 90% to 70% correct. Group performance for vowel identification in each octave is presented in Fig. 4 as a function of the number of cycles. Performance for vowel identification in octaves 1–3 was at ceiling, whereas that for vowel identification in octave 4 improved as the number of cycles increased. Note and octave value interacted with the number of cycles ($F_{54,324}=6.41$, $p=0.04$). This interaction occurred because performance at the highest notes improved as the number of cycles increased.

The interactions reported for vowel identification were especially evident for the vowel /ɜ/. This was shown by a significant interaction between octave, note, and vowel ($F_{27,162}=8.41$, $p<0.001$), and the number of cycles, octave, note, and vowel ($F_{162,972}=1.70$, $p<0.001$).

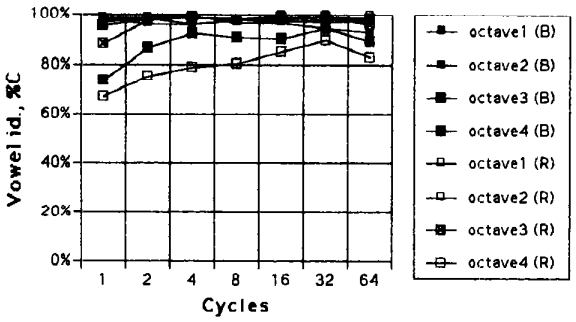

FIG. 4. Psychometric functions for vowel identification as a function of the number of cycles in octaves 1–4 for blocked and random stimulus presentation.

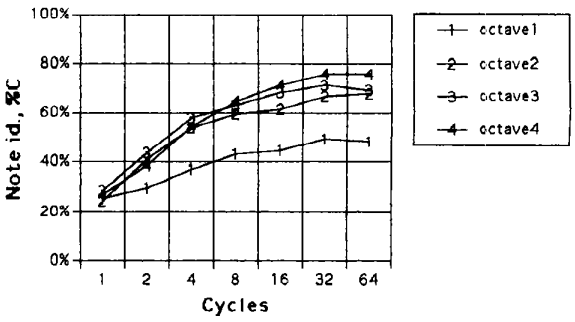

FIG. 6. Psychometric functions for note identification as a function of the number of cycles in octaves 1–4. Performance has been collapsed over blocked and random stimulus presentation.

B. Octave identification

Vowel quality affected octave identification and the interaction was significant for both versions of the experiment, blocked and randomized ($F_{9,54}=7.87$, $p<0.001$). An examination of the confusion matrix showed that octave errors for vowels /a/ and /i/ were more likely to be at a higher octave, whereas octave errors with /u/ vowels were more likely to be at a lower octave. This is presumably because there is relatively more low-frequency energy in /u/ (Fig. 2).

There was also an interaction between vowel, octave, and number of cycles ($F_{54,324}=2.10$, $p<0.001$). It was more apparent in the randomized version of the experiment, so these data were analyzed separately. The total number of "positive" and "negative" octave responses, for each vowel at each stimulus duration, is presented in Fig. 5. The "+" symbols indicate that the perceived octave was higher than the stimulus octave, whereas the "−" symbols indicate that perceived octave was lower than the stimulus octave. So if an octave 2 stimulus attracted an "octave 1" response, it was coded as negative, whereas "octave 3" and "octave 4" responses were coded as positive. The abscissa is the number of cycles; the ordinate is the percentage of positive or negative responses for all listeners combined. The figure shows

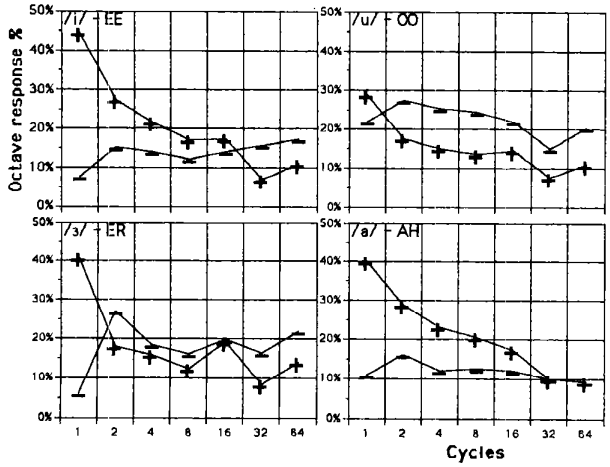

FIG. 5. Higher (+) and lower (−) octave responses as a function of cycles for each vowel. For details refer to text.

that for /i/ and /a/ there are more positive octave errors and they are associated with the briefer stimuli. For /ɜ/, there are fewer octave errors and they are almost evenly distributed between positive and negative. For /u/ there are more negative octave errors and they occur at both shorter and longer durations. If we note (a) that the vowels /i/ and /a/ have relatively more high-frequency energy, while the vowel /u/ has relatively more low-frequency energy, and (b) that low-frequency auditory filters are narrower and therefore slower than higher-frequency filters, then it might be possible to explain this interaction as follows: When the stimulus is short, the internal representation is not well defined, and in this case, the center of gravity of the spectrum of the sound influences the perceived octave. As the duration of the vowel increases, the internal spectrum becomes better defined and the number of octave errors decreases. The effect extends to longer durations for vowels with predominantly low-frequency energy because longer durations are required to achieve the same degree of definition in the internal spectrum.

There was one other interaction: the octave in which the sound was presented affected average performance; mean percent correct for octaves 1–4 was 75, 63, 67, 79, respectively. A *post hoc* analysis showed that octaves 1 and 4 together were better identified than octaves 2 and 3 together ($F_{3,18}=12.10$, $p<0.001$). The observation that the end points of the stimulus range are better identified than the midpoints is characteristic of judgments based on a continuous dimension, as opposed to those based on categories.

C. Note identification

Notes presented at the lowest octave were not as well identified as notes in the higher octaves ($F_{3,18}=34.48$, $p<0.001$). Performance at octave 1 was at 40% correct compared with 54%, 58%, and 58% correct for octaves 2–4. There was an octave by cycle interaction ($F_{18,108}=3.38$, $p<0.001$) which occurred because note identification at the lowest octave did not improve as rapidly as note identification in the higher octaves. This may be seen in Fig. 6 which shows note identification in octaves 1–4 as a function of the number of cycles. The effect is probably due to the fact that these notes are close to the lower limit of pitch perception.

The finding that note identification improves with number of cycles is consistent with the data of Ritsma *et al.* (1965). They studied the pitch matching of complex tones which ranged in $f0$ from 71 to 560 Hz and found that the standard deviation of pitch matching decreased as a duration increased. When their data are presented as a function of the number of cycles, they show that the standard deviation of pitch matches decreased from 3% at one cycle to 0.5% between six and eight cycles.

Vowel quality was also found to interact with note identification. The interaction between vowel and note ($F_{9,54}=6.62$, $p<0.001$) was caused by the note of the /u/ stimuli being better identified for the note C (63% correct) than for the other three vowels (/a/ 56%, /i/ 52%, /ɜ/ 55% correct). This was confirmed by multiple comparisons ($p<0.05$). There was also a significant interaction between vowel, note, and duration ($F_{54,324}=2.22$, $p<0.001$), but the confusion matrices did not reveal the compelling effects observed with octave identification.

Direct evidence for the effect of vowel quality on pitch may be found in the studies of Chuang and Wang (1978) and of Stoll (1984). Chuang and Wang presented listeners with pairs of vowels, one of which was always an /a/ with a fundamental of 100 Hz. The other vowel was /ɜ/, /i/, or /u/ and its fundamental was varied in the region of 100 Hz over trials. Listeners were asked whether they heard a difference in pitch between the two vowels. They found that the matching pitches for /ɜ/, /i/, and /u/ were lower than that of /a/ by 0.5, 1.3, and 2.8 Hz, respectively. They attributed the pitch effects to the influence of the first formant; all three vowels have first formant frequencies which are lower than that of the vowel /a/. Stoll subsequently confirmed the finding, and suggested that these spectral effects were due to the effects of harmonics in the dominant region (Ritsma, 1967; Plomp, 1967). As the first formant lies in the dominant region, it appears that the relative level of harmonics associated with the first formant has a small, but statistically significant, effect on pitch.

III. DISCUSSION

The contrast between the experimental results at the longer and shorter durations would appear to cast considerable doubt on the concept of pitch and timbre extraction presented in recent computational models of complex sound segregation (Scheffers, 1983; Assmann and Summerfield, 1990; Meddis and Hewitt, 1992). These models assume that the auditory system can extract accurate, absolute pitch values from brief segments of multisource sounds, and use these pitch values to direct the extraction of timbre information from the complex neural activity patterns flowing from the cochlea in response to these sounds. The experimental results show that performance on a four-category vowel identification task is near ceiling even with one cycle of the sound. In contrast, performance on a four-category pitch-chroma identification task is at chance levels with one cycle of the sound and only rises to about 65% correct as the number of cycles increases to 64. Performance on the easier pitch task, octave identification, falls midway between vowel and pitch-chroma identification.

It is not surprising to find that vowel identification is an easier task than pitch-chroma identification. However, it is difficult to understand why it is a more difficult task if absolute pitch values are extracted prior to timbre identification as suggested in the computational models of vowel segregation. Vowel identification is a highly overlearned task, but if absolute pitch values are required for timbre extraction, then we might expect that note identification would be equally overlearned, or at least better than the near chance performance we observe with small numbers of cycles. In all likelihood, auditory extraction of pitch and timbre information is much more complicated than it is portrayed in existing computational models of sound segregation. It could be that pitch-chroma values extracted from brief stimuli have sufficient accuracy to support good pitch-chroma performance, but there is some limitation in the more central, note-naming process that requires the pitch values to be available for a longer duration. But in the absence of explicit models of this form, the argument will not be pursued.

The advantage of overlearning on the difference between vowel and pitch-chroma identification can be estimated from a comparison of the current results with those reported in Robinson and Patterson (1995). They performed a similar four-category identification experiment to that reported in this paper but using notes from musical instruments rather than vowels. The results for pitch-chroma identification and octave identification are similar to those reported in the current study. Specifically, pitch-chroma performance rises from chance with one-cycle stimuli to about 75% correct with 64-cycle stimuli. The instrument identification results are like the vowel identification results insofar as performance with short duration stimuli is better than for octave and pitch-chroma identification and the instrument identification performance does not improve with increasing numbers of cycles. However, the level of performance for instrument identification is about 70% correct, whereas it is about 95% correct for vowel identification. Thus, at the longest durations in the instrument experiment, performance on the pitch-chroma identification task is equal to, or perhaps a little better than, performance on the instrument identification task. But at the shorter durations, performance on the timbre task is decidedly superior to that on the pitch-chroma task. Performance on the octave task falls between that for instruments and notes. Thus there is a clear advantage for vowel identification, but performance on the timbre tasks in both experiments is reliably above chance and near its asymptotic level at the shortest durations whereas pitch-chroma identification is at chance with one cycle of the sound and does not reach asymptotic levels until about 32 cycles of the sound.

It might also be possible to pursue the original idea, that pitch is used to assist timbre extraction, within the framework of the more extensive virtual-pitch model proposed by Terhardt (1987). In this model, the processing of pitch and timbre begins with the extraction of individual spectral pitches which are then combined to determine the virtual pitch and the timbre of the sound. The accuracy of spectral pitch values improves with duration at the start of a sound in any model, so it may well be that these initial estimates have sufficient accuracy to specify the timbre of the sound for the

instrument identification task, and at the same time, insufficient accuracy to specify the virtual pitch of the sound for the pitch-chroma task. In this case, the influence of pitch on timbre extraction would be limited to the influence of spectral pitch values, since the virtual pitch is assumed to require the resolution associated with longer duration stimuli. So this would not solve the problem for the computational sound segregation models, because they use the analog of virtual pitch to guide vowel segregation. Nevertheless, the example illustrates how more complicated auditory models might deal with the limitation that the pitch-chroma data would appear to impose.

Finally, it is worth reversing the perspective for a moment and noting that the results from the current experiment and the experiment of Robinson and Patterson (1995) might be used to argue that timbre information is used to assist pitch extraction. This will not be pursued in the current paper, but it is a more reasonable hypothesis than the reverse.

IV. CONCLUSIONS

Vowel identification is about 90% correct with one cycle of the stimulus, and rises to ceiling performance at two cycles. Octave identification is above chance but below ceiling performance with one cycle of the stimulus (about 53% correct). Note identification is at chance performance with one cycle of the stimulus, and rises steadily as the number of cycles increases to about 65% correct at 64 cycles.

There were several interactions: Vowel quality affected both octave and note identification with the greatest effect at short durations. Fundamental frequency affected vowel identification for the highest notes, and the greatest effect was found at longer durations.

This pattern of results indicates that pitch is not required for timbre perception, and it is more consistent with theories of auditory perception where timbre and pitch are processed in parallel.

ACKNOWLEDGMENT

The work was supported by a grant from the UK Defense Research Agency, Farnborough (Project AAM HAP).

Assman, P. F., and Summerfield, A. Q. (**1989**). "Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency," J. Acoust. Soc. Am. **85**, 327–338.

Assman, P. F., and Summerfield, A. Q. (**1990**). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680–697.

Brown, G. J., and Cooke, M. (**1994**). "Computational auditory scene analysis," Comput. Speech Lang. **8**, 297–336.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech with one, and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

Cheveigne, A. de (**1993**). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. **93**, 3271–3290.

Chuang, C.-K., and Wang, W. S.-Y. (**1978**). "Psychophysical pitch biases related to vowel quality, intensity difference, and sequential order," J. Acoust. Soc. Am. **64**, 1004–1014.

Duifhuis, H., Willems, L. F., and Sluyter, R. J. (**1982**). "Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception," J. Acoust. Soc. Am. **71**, 1568–1580.

Fant, C. G. M. (**1962**). "Descriptive analysis of the acoustic aspects of speech," Logos **5**, 3–17.

Goldstein, J. L. (**1973**). "An optimum processor theory for the central formation of the pitch of complex tones," J. Acoust. Soc. Am. **54**, 1496–1516.

Gray, G. W. (**1942**). "Phonemic microtomy: The minimum duration of perceptible speech sounds," Speech Monogr. **9**, 75–90.

Henning, G. B. (**1967**). "Frequency discrimination in noise," J. Acoust. Soc. Am. **41**, 774–777.

Klatt, D. H. (**1979**). "Speech perception: A model of acoustic–phonetic analysis and lexical access," J. Phon. **7**, 279–312.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 279–312.

Licklider, J. C. R. (**1951**). "A duplex theory of pitch perception," Experientia **7**, 128–133.

Meddis, R., and Hewitt, M. J. (**1992**). "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. **91**, 233–245.

Patterson, R. D. (**1987**). "A pulse ribbon model of peripheral auditory processing," in *Auditory Processing of Complex Sounds*, edited by W. A. Yost and C. S. Watson (Erlbaum, Hillsdale, NJ), pp. 167–179.

Patterson, R. D. (**1990**). "The tone height of multi-harmonic sounds," Music Percept. **8**, 203–214.

Patterson, R. D., and Hirahara, T. (**1989**). "HMM speech recognition using DFT and auditory spectrograms," ATR HIP Tech. Rep. 10.23.89.

Patterson, R. D., Milroy, R., and Allerhand, M. (**1993**). "What is the octave of a harmonically rich note?," in *Proceedings of the 2nd International Conference on Music and the Cognitive Sciences*, edited by I. Cross (Harwood, London), pp. 69–81.

Patterson, R. D., Peters, R. W., and Milroy, R. (**1983**). "Threshold duration for melodic pitch," in *Hearing—Physiological Bases and Psychophysics*, edited by R. Klinke and W. Hartmann (Springer-Verlag, Berlin), pp. 321–325.

Plomp, R. (**1967**). "Pitch of complex tones," J. Acoust. Soc. Am. **41**, 1526–1533.

Ritsma, R. J. (**1967**). "Frequencies dominant in the perception of the pitch of complex sounds," J. Acoust. Soc. Am. **42**, 191–198.

Ritsma, R. J., Cardozo, B. L., Domburg, G., and Neelen, J. J. M. (**1965**). "The buildup of the pitch percept," IPO Rep. **1**, 12–15.

Robinson, K. L. (**1993**). "Studies in timbre and pitch," Ph.D. thesis, Cambridge University.

Robinson, K. L., and Patterson, R. D. (**1995**). "The duration required to identify the instrument, the octave, or the pitch-chroma of a musical note," Music Percept. **13**, 1–15.

Scheffers, M. T. M. (**1983**). "Sifting vowels: Auditory pitch analysis and sound segregation," Ph.D. thesis, University of Groningen.

Stoll, G. (**1984**). "Pitch of vowels: Experimental and theoretical investigation of its dependence on vowel quality," Speech Commun. **3**, 137–150.

Stumpf, C. (**1890**). *Tonpsychologie* (Hirzel, Leipzig). Republished in 1965 (Knef/Bonset, Hilversum/Amsterdam).

Suen, C. Y., and Beddoes, M. P. (**1972**). "Discrimination of vowel sounds of very short duration," Percept. Psychophys. **11**, 417–419.

Terhardt, E. (**1987**). "Psychophysics of audio signal processing and the role of pitch in speech," in *The Psychophysics of Speech Perception*, edited by M. E. H. Schouten (Nijhoff, Leiden), pp. 271–283.