Contributions of temporal encodings of voicing, voicelessness, fundamental frequency, and amplitude variation to audio-visual and auditory speech perception

Andrew Faulkner and Stuart Rosen
Department of Phonetics and Linguistics, University College London, 4 Stephenson Way, London NW1 2HE, United Kingdom

(Received 7 December 1998; revised 17 May 1999; accepted 11 June 1999)

Auditory and audio-visual speech perception was investigated using auditory signals of invariant spectral envelope that temporally encoded the presence of voiced and voiceless excitation, variations in amplitude envelope and F_0. In experiment 1, the contribution of the timing of voicing was compared in consonant identification to the additional effects of variations in F_0 and the amplitude of voiced speech. In audio-visual conditions only, amplitude variation slightly increased accuracy globally and for manner features. F_0 variation slightly increased overall accuracy and manner perception in auditory and audio-visual conditions. Experiment 2 examined consonant information derived from the presence and amplitude variation of voiceless speech in addition to that from voicing, F_0, and voiced speech amplitude. Binary indication of voiceless excitation improved accuracy overall and for voicing and manner. The amplitude variation of voiceless speech produced only a small increment in place of articulation scores. A final experiment examined audio-visual sentence perception using encodings of voiceless excitation and amplitude variation added to a signal representing voicing and F_0. There was a contribution of amplitude variation to sentence perception, but not of voiceless excitation. The timing of voiced and voiceless excitation appears to be the major temporal cues to consonant identity. © *1999 Acoustical Society of America.* [S0001-4966(99)01410-1]

PACS numbers: 43.71.Es, 43.71.Ma, 43.66.Ts [JMH]

INTRODUCTION

The perceptual role of temporal structure in speech has until recently been rather neglected in comparison to spectral structure (Rosen, 1992; van Tasell *et al.*, 1987, 1992). Temporal speech information is likely to have special significance for hearing-impaired listeners, for users of cochlear implants, and more generally in listening conditions where noise masks spectral structure. Both electro-cochlear stimulation (Shannon, 1993) and acoustic stimulation to even profoundly hearing-impaired listeners (Faulkner *et al.*, 1992; Rosen *et al.*, 1990) are typically perceived with sufficient temporal resolution for temporal speech information to be processed. Cochlear hearing impairment of moderate or greater degree is typically associated with significant degradation of frequency selectivity (e.g., Moore, 1996). At frequencies where the hearing loss is profound, frequency selectivity can be completely absent (Faulkner *et al.*, 1990). The partial loss of spectral detail appears to have rather slight effects on the perception of speech at least in quiet (Baer and Moore, 1993; Shannon *et al.*, 1995). This result suggests that temporal cues and changes of gross spectral shape may be of greater significance in speech perception than has often been thought.

The notion that segmental and supra-segmental speech information needs to be clearly represented by speech perceptual prostheses is well-established. The present studies are largely concerned with the contributions from different aspects of temporal speech information to consonant perception in conditions where spectral structure is eliminated. They have both a theoretical aim in relation to accounts of speech temporal structure (e.g., Rosen, 1992) and a practical aim in identifying those aspects of temporal structure that need to be effectively transmitted through a hearing aid or cochlear implant.

A. Relation to previous studies

Many studies have examined the contribution of temporally coded cues to the audio-visual perception of connected speech (e.g., Breeuwer and Plomp, 1986; Grant *et al.*, 1985, 1991; Risberg and Lubker, 1978; Rosen *et al.*, 1981; Waldstein and Boothroyd, 1994). However, studies of sentence-level materials provide little insight into the phonetic information conveyed by temporal cues. Segmental perception from temporal cues, which can be analyzed to be informative of temporal contributions to phonetic features, has received much less attention.

Studies of consonant identification from temporally coded auditory information have rarely included both auditory and audio-visual conditions. Here we have studied consonant identification in both auditory and audio-visual conditions and also from purely visual input, so that the contributions of temporal structure can be observed both with and without partially complementary visual cues. Moreover, such studies of consonant perception have not sought to isolate components of temporal structure in relation to the acoustic-phonetic structures of speech (Baker and Rosen, 1994; Breeuwer and Plomp, 1986; Rosen, 1992; van Tasell *et al.*, 1987, 1992). In particular, the role of periodic and aperiodic excitation has largely been ignored. Furthermore,

in consonant identification, amplitude variation has never been dissociated from simple duration cues, although this has been done for connected speech (Breeuwer and Plomp, 1986; Grant *et al.*, 1985).

B. The temporal structure of speech

Temporal information has several component elements which have been associated with both segmental and supra-segmental speech information. Rosen (1992) proposed a three-way classification of temporal components by rate. This is linked to Fourcin's (1977, 1990) analysis of speech in terms of acoustic speech patterns that relate to putatively significant productive and receptive features. Low rate temporal information (below about 50 Hz) conveys the presence or absence of acoustic excitation and the amplitude variation of periodically and aperiodically excited speech. This factor, termed *envelope* by Rosen (1992), is correlated with the effort underlying speech excitation, the degree of constriction of the vocal tract, and velar closure. Temporal information at rates between 50 and a few hundred Hz represents the periodicity or aperiodicity of excitation, and voice fundamental frequency (F_0). This factor was termed *periodicity* in Rosen's classification. A still higher rate component of temporal speech information is that contained in the *fine structure* of the speech pressure waveform. This contains information related to vocal tract resonances. There is no strong evidence that speech spectral structure can be perceived from temporal fine structure, and it is not considered in the present study.

These classes of information, although broadly assigned to temporal rate ranges, cannot be properly separated by modulation rate. For example, both F_0 and $F1$ may affect temporal modulations in the 200–500 Hz region, and at burst release, the onset of voiceless excitation can be extremely rapid. This conceptual framework nevertheless allows a broad classification of the production-related acoustic pattern information that exists over the speech modulation rate spectrum.

Temporal information at low and moderate rates has particular significance in speechreading, because it is largely uncorrelated with visible oral correlates of place of articulation, and is hence perceptually complementary to speechreading (Summerfield, 1987). Spectral speech information, conversely, is broadly correlated with visible information from speechreading.

C. Phonetic information in temporal structure

1. Patterning of excitation

Much of the voicing and manner of articulation information in speech has acoustic correlates in the patterning in time of quasi-periodic laryngeal excitation, silence, and aperiodic excitation. Hence it would be expected that temporal cues can signal such information from the timing of the presence and absence of excitation and from its periodicity or aperiodicity. The extent to which the timing of voiced excitation contributes to consonant identification is examined in experiment 1. The additional consonantal information available from the timing of voiceless excitation is examined in experiment 2. The significance of a cue to the timing of voiceless excitation in connected speech is addressed in experiment 3.

2. Amplitude variation

The degree of oral constriction and the opening of the nasal tract both affect the amplitude envelope of voiced speech. Hence, the variation of speech amplitude might be expected to provide cues to manner of articulation (e.g., Shinn and Blumstein, 1984). The envelope of voiceless speech is influenced both in its duration and its amplitude by the place of articulation of voiceless fricatives and voiceless plosives. Speech amplitude is also affected by the amplitude variations of both periodic and aperiodic excitation that follow from variation in sub-glottal pressure. Such variation is broadly associated with prosodic features and is not expected to be significant in consonant identity. Experiment 1 examines the role of the amplitude variation of voiced speech in consonant identification. Experiment 2 includes conditions which allow an assessment of the contribution of voiceless speech amplitude variation in consonant identification. Experiment 3 addresses the role of amplitude variation in the audio-visual perception of connected speech.

3. Fundamental frequency

The perceptual significance of voice F_0 is primarily supra-segmental. However, F_0 is affected by changes in the acoustic impedance of the vocal tract resulting from differing degrees of occlusion during consonant articulation (Fant, 1973, p. 80). Voiced consonants may thus contain F_0 variation that carries some salient information related to manner of articulation. Experiment 1 examines this hypothesis.

D. Interpretation of amplitude envelope

Investigators have used a number of different definitions and measurement methods for amplitude envelope. The perceptual contribution of amplitude envelope depends both on the speech frequencies over which it is measured and on the degree of smoothing applied to the extracted envelope (e.g., Grant *et al.*, 1991). Where the amplitude envelope smoothing filter extends into the voice F_0 range, typical methods of envelope extraction lead to the periodicity of F_0 being included in the envelope signal. In order to distinguish periodicity from other components of the amplitude envelope, we have used an envelope smoothing filter that is below the F_0 range, and have derived periodicity information from larynx activity.

When amplitude envelope is measured directly from the speech signal, the choice of bandwidth from which amplitude envelope is measured determines whether voiceless speech energy is included. Rather than take this purely acoustic approach to separate voiced and voiceless excitation, we have used a reference signal from laryngeal activity to distinguish voiced and voiceless excitation. Henceforth, we use the term amplitude variation to refer to low rate amplitude fluctuations up to around 50 Hz. For voiced speech, amplitude envelope has been derived from frequencies up to

3 kHz. For voiceless speech components, amplitude envelope was measured here over the frequency range 3–10 kHz.

I. TEMPORAL INFORMATION IN CONSONANT IDENTIFICATION

A. Experiment 1: Segmental information from voicing pattern, amplitude variation, and F_0 variation

This first study examined the contributions to consonant identification of the gross timing, amplitude variation, and F_0 variation of voiced components of speech. Because of the segmental correlates of these three temporal factors, each may be expected to have a possible role in cueing consonantal features and identity. This experiment is designed to evaluate their relative significance both in audio-visual and purely auditory perception using auditory signals with an invariant spectral envelope.

1. Method

a. Conditions. The contributions of these temporal components in auditory and audio-visual consonant identification are assessed here by differences in performance across nine different presentation conditions. Four auditory signals were used, both with and without visual lipreading information. A purely visual lipreading condition, **L**, was also included. The gross timing of voicing was represented by signal **V**, a pulse-train of fixed frequency and amplitude, which was present only during vocal fold vibration. Signal **V(A)** differed from signal **V** in carrying amplitude envelope variation, derived from the original speech. F_0 variation without amplitude variation was present in signal **F**, where a fixed-amplitude pulse-train during voicing followed the speaker's F_0. Finally, signal **F(A)** varied in both frequency and amplitude according to F_0 and the amplitude envelope of voiced speech, as well as indicating the timing of voicing.

b. Stimulus processing. Fundamental frequency and the timing of laryngeal excitation were derived from an electrolaryngograph signal that had been recorded with the speech signal. This processing made use of a Voiscope™ (Laryngograph Ltd.[1]) which generated a brief logic pulse at each vocal fold closure, and a flip-flop logic signal that indicated voicing. The voicing detector was used to gate a pulse-train signal. For voiced speech the rate of this pulse train was fixed at 200 Hz for signals **V** and **V(A)**. For signals **F** and **F(A)**, the pulse rate was controlled cycle-by-cycle from the signal generated at each larynx closure so as to follow the voice fundamental frequency.

Amplitude envelope was derived by full-wave rectifying the speech signal after 3-kHz low-pass filtering (Kemo VBF/3: 48 dB/octave) and then by smoothing the rectifier output with a low-pass filter having its −3-dB point at 22.5 Hz, and subsequent attenuation of 18 dB at 50 Hz and 53 dB at 80 Hz (Kemo VBF 8[2]).

Finally, the pulse signal was multiplied by the amplitude envelope and then low-pass filtered at 400 Hz (Kemo VBF/14: 18 dB/octave). The processing of voicing and F_0 involved an audio delay of one period. The amplitude envelope measure involved a delay of 19.5 ms arising from the impulse response of the smoothing filter. Speech input and output signals illustrating the result of this processing are shown in Fig. 1.

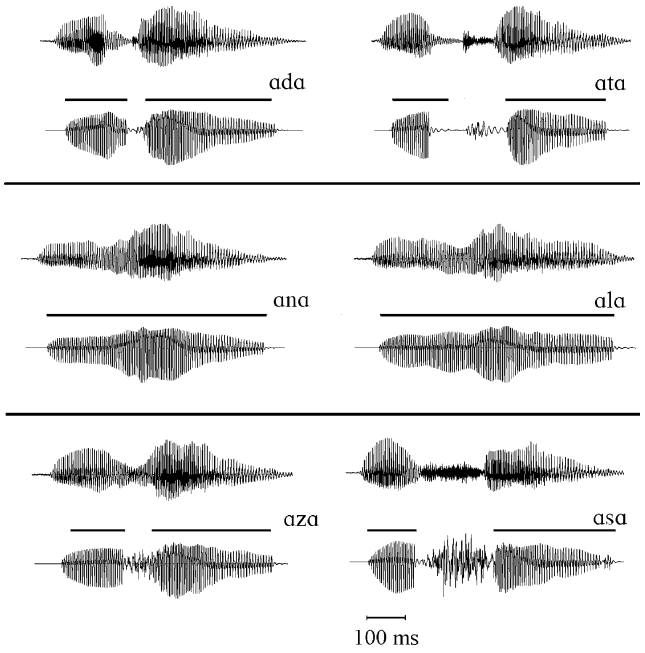

FIG. 1. Speech (upper) and stimulus (lower) waveforms for six alveolar VCVs. The stimulus waveform is for condition **F(A)+N(A)** of experiment 2. Signal **F(A)** used in experiment 1 is similar except that the signal level is zero where voicing was not detected. The lines above the stimulus waveforms mark the timing of voiced excitation as detected from the Laryngograph signal.

Speech materials comprised each of the 24 English consonants between the vowel /a/. Five video-recorded lists from a female speaker of standard southern British English were used, comprising in total ten distinct tokens of each consonant, with two in each of the lists. Each list used a different pseudo-random ordering. The stimuli were originally recorded onto Hi Band U-matic video, with the speech signal and an electrolaryngograph signal on the two FM audio channels. These recordings were then edited and copied to S-VHS using the FM audio tracks. The stimuli presented in the experiment itself were copied to VHS video tapes, using the FM audio tracks to record all five lists with each of the auditory signals.

2. Subjects

Five normally hearing subjects aged between 18 and 30 with normal or corrected-to-normal vision took part. All were native speakers of British English. They were paid for their participation.

3. Procedure

All the experiments reported here took place in an acoustically isolated room. The video image was presented on a 36-cm Panasonic color monitor. The audio signal was presented free-field at approximately 65 dBA through either a Rogers A75 amplifier and Rogers LS3/5a loudspeaker or a

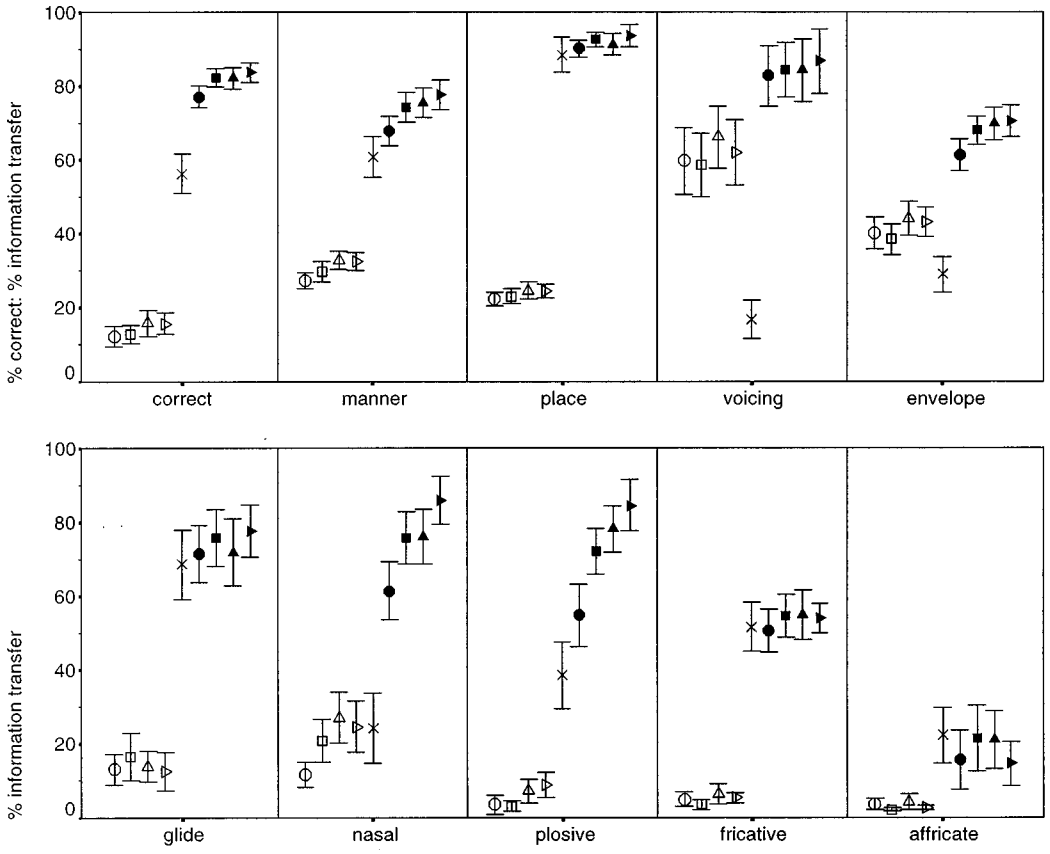

FIG. 2. Upper panel: percent correct and percent information transfer for manner, place, voicing, and envelope features. Lower panel: percent information transfer for binary manner features. Data from experiment 1, five subjects. Conditions are shown as **V**: ○; **V(A)**: □; **F**: △; **F(A)**: ▷: **L**: ×. Audio-visual conditions have the same symbol as the corresponding auditory conditions, but filled. Error bars show ±2 standard errors over 30 data points (subjects ×sessions).

Yamaha P2050 amplifier and a QUAD PRO-63 loudspeaker. Subjects sat approximately 1.5 m from the monitor and loudspeaker.

The first of the five lists of consonants was reserved for training, which was provided before the first test list in each condition. Training was performed by the experimenter providing verbal feedback to subjects after they had attempted to identify each one of the 48 training stimuli. The eight audio and audio-visual conditions were presented in blocks of four test lists for which the audio signal remained the same. Five such blocks were presented in total, interleaved with five blocks of two lists in condition **L**, with the blocks in a randomized order. The order of auditory and audio-visual conditions within blocks was counterbalanced, and the ordering of test lists in each block was randomized to minimize learning of the stimulus order.

Subjects responded in writing on a printed response sheet during a blank time interval following each stimulus. In total, 10 lists of 48 consonants were administered to each subject in each condition.

4. Results

Consonant identification data from each test list were analyzed to give overall proportion correct scores, together with information transfer measures (Miller and Nicely, 1955) for the binary voicing feature, for place of articulation (distinguishing bilabial, labiodental, dental, alveolar, palatal, velar, and pharyngeal place), for manner (plosive, affricate, fricative, nasal, or glide), and for the modified envelope feature defined by van Tasell *et al.* (1992). The envelope feature distinguishes four classes of consonant: voiceless plosives; voiced plosives and fricatives; voiceless fricatives; nasals and glides. Subjects showed improving performance over the early sessions. The analyses presented here are based on the last six lists in each condition, which showed more stable levels of performance over sessions.

Preliminary repeated measures analyses of variance for the overall accuracy and feature measures used factors of condition, test session, and subject. These revealed significant effects of condition, session, and subject for all measures. There were no significant interactions between session and condition and the session factor was henceforth ignored. A second series of repeated measures ANOVAs of overall correct and feature scores that excluded condition **L** treated the experiment as a 2×2×2 factorial design, with factors of amplitude variation, F_0 variation, and visual information, in addition to subject. Between session variability was included in the residual error term.[3]

Mean scores in each condition are shown in Fig. 2. Auditory only scores were always rather low except for voicing and envelope features. Visual information had a highly significant effect on all of the measures. F_0 variation was a significant main effect for overall accuracy $[F(1,8) = 15.9, p = 0.004]$, for manner $[F(1,12) = 20.3, p = 0.001]$,

voicing [$F(1,228)=4.0, p=0.047$], and envelope information [$F(1,228)=20.1, p=0.001$]. The effects of F_0 variation did not interact with any other factor.

Amplitude variation showed a significant main effect for both overall accuracy [$F(1,216)=5.17, p=0.024$] and for manner information [$F(1,12)=6.41, p=0.026$]. However, there was a significant interaction of amplitude variation with the presence of visual information for overall accuracy [$F(1,216)=4.18, p=0.042$] and the envelope feature [$F(1,228)=5.16, p=0.024$]. This interaction also came close to significance for manner information [$F(1,212)=3.34, p=0.069$]. Hence, further analyses were carried to test the effect of amplitude variation for auditory and audio-visual conditions separately. In auditory conditions, the effect of amplitude variation was never significant for these measures (p was always greater than 0.3). However, in audio-visual conditions, these analyses were able to reveal small but significant effects of amplitude variation on overall accuracy [$F(1,112)=9.4, p=0.003$], manner [$F(1,112)=9.4, p=0.003$], and envelope information [$F(1,112)=4.84, p=0.030$].

The effects of F_0 and amplitude variation as they act on consonant manner contrasts have been examined in more detail by computing information transfer scores for binary manner features of nasal, plosive, fricative, affricate, and glide. These are included in Fig. 2. Affricate, fricative, and glide features proved to be affected only by the presence or absence of visual information. Nasal and plosive features were affected by visual information and also by amplitude and F_0 variations. The presence of F_0 variation significantly increased nasal information across auditory and audio-visual conditions [$F(1,4)=11.1, p=0.029$], while F_0 variation showed an interaction with visual information for the plosive feature. Further analysis showed a significant effect of F_0 on plosive information in audio-visual conditions only [$F(1,112)=40.3, p<0.001$]. Amplitude variation showed significant interactions with visual information for both nasal and plosive features. In auditory conditions, there was no significant effect of amplitude variation, while in audio-visual conditions, there were significant amplitude effects for both nasal [$F(1,112)=11.7, p<0.001$] and plosive information [$F(1,112)=17.9, p<0.001$].

5. Discussion

The results indicate that the bulk of the auditory information in these stimuli comes from the timing of voiced excitation. F_0 variation had a significant, but modest effect, increasing overall accuracy by just over 3%. This followed from the improved perception of nasal and plosive manners. The amplitude variation of voiced speech at rates of up to 50 Hz led to significant increments in accuracy (again of about 3%) and manner information in audio-visual conditions, but had no effect in purely auditory conditions.

B. Experiment 2: Segmental information from voiceless excitation

A second study addressed the role of voiceless frication and the amplitude envelope of voiceless speech. Here there were three different acoustic signals each presented with and without lip-reading.

1. Conditions

The **F(A)** signal used in experiment 1, encoding timing of voicing, F_0 and voiced speech amplitude, was retained as a reference condition and was compared to two new signals that represented voiceless excitation. In condition **F(A)+N**, the **F(A)** signal was combined with a fixed-level aperiodic noise that was present during purely voiceless excitation. The contribution of amplitude variation in voiceless speech was examined by comparing **F(A)+N** with **F(A)+N(A)**, in which the noise component was modulated by the amplitude envelope of voiceless speech.

2. Stimuli

Voiced speech was processed as before. Voiceless excitation was detected by a threshold applied to the output of a spectral balance circuit that compared the amount of energy above and below 3 kHz in the speech signal. Voiceless excitation was represented by a 400-Hz low-pass filtered noise, only in the absence of voicing as detected from the recorded electrolaryngograph signal. Amplitude envelope for voiceless speech was derived by full-wave rectifying the speech signal after a 3-kHz high-pass filter (Kemo VBF/3: 48 dB/octave), and smoothing the result as for voiced speech using a low-pass filter with its -3-dB point at 22.5 Hz (Kemo VBF/8, as experiment 1). The pulse and noise signals were added together after multiplication by the separately extracted amplitude envelopes of voiced and voiceless speech.

3. Subjects and procedure

Five new subjects with normal hearing and normal or corrected to normal vision took part and were paid for their services. Procedures were the same as those for experiment 1 and again ten test lists of 48 stimuli were presented in each condition.

4. Results

Data from the last six test lists in each condition only were analyzed. Results are shown in Fig. 3. As for experiment 1, ANOVAs by condition, session, and subject were carried out for each measure. Test session had a significant effect on most scores, but there were no significant interactions between condition and session. The main analyses used a partial factorial design, with factors of visual information, a cue to the presence of voiceless excitation [comparing signals **F(A)** and **F(A)+N**], and the encoding of amplitude variation for voiceless speech [comparing signals **F(A)+N** and **F(A)+N(A)**]. These revealed highly significant ($p<0.001$) main effects of visual information for overall accuracy and for each of the features of manner, place, voicing, and envelope. Cueing the presence of voiceless excitation led to significant increases in all of these measures except for place information; overall accuracy, $F(1,116)=30.7$, $p<0.001$; manner, $F(1,162)=84.4$, $p<0.001$; voicing, $F(1,17.2)=11.35$, $p=0.004$; envelope, $F(1,4)=55.67$, $p=0.002$.

The encoding of amplitude variation for voiceless speech showed a main effect only for place information, where there was a small but significant increase in scores in

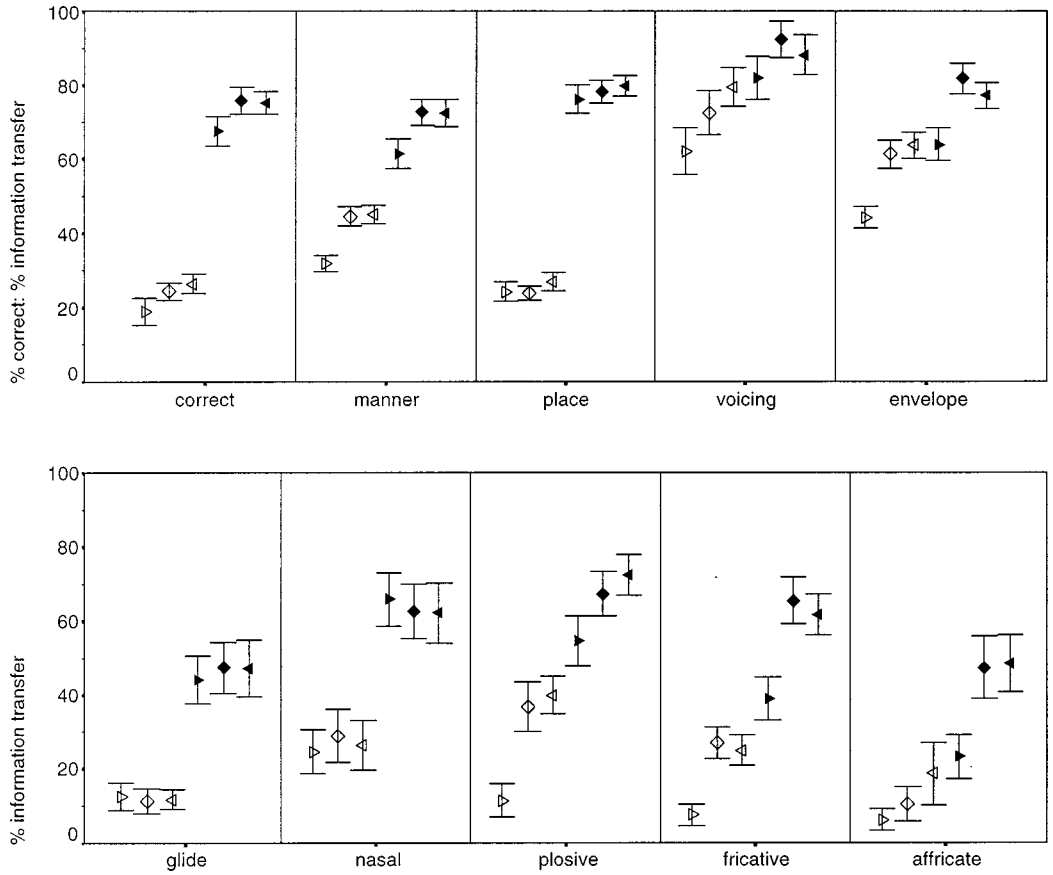

FIG. 3. Upper panel: percent correct and percent information transfer for manner, place, voicing, and envelope features. Lower panel: percent information transfer for binary manner features. Data from experiment 2, five subjects. Conditions are shown as **F(A)**: ▷; **F(A)+N**: ◇; **F(A)+N(A)**: ◁. Audio-visual conditions have the same symbol as the corresponding auditory conditions, but filled. Error bars show ±2 standard errors of 30 data points.

both auditory and audio-visual conditions in the presence of this information [$F(1,158)=4.75, p=0.031$]. There were significant interactions between voiceless speech amplitude variation and visual information for voicing [$F(1,158)=5.93$, $p=0.016$] and envelope features [$F(1,158)=5.66$, $p=0.019$]. Both were crossover interactions, reflecting an increase in scores with amplitude variation in auditory conditions but a decrease in audio-visual conditions.

Binary manner features (glide, nasal, plosive, fricative, and affricate: see Fig. 3) all showed highly significant increases with the presence of visual information ($p<0.006$). As in experiment 1, glide and nasal features were unaffected by differences in the auditory signals. Encoding the presence of voiceless excitation led to significant increases in information transfer for plosive [$F(1,166)=54.6, p<0.001$], fricative [$F(1,10.32)=56.8, p=0.001$], and affricate features [$F(1,158)=19.3, p<0.001$]. For plosive and affricate features, this factor also showed a significant interaction with the presence of visual cues. For the plosive feature, this interaction reflects a somewhat larger increase in scores in auditory conditions than audio-visually. In the case of the affricate feature, however, there was little effect in auditory conditions of the voiceless cue, but a large effect audio-visually. The only binary manner feature affected by the amplitude variation of voiceless speech was the plosive feature [$F(1,17.6)=5.61, p=0.029$].

5. Discussion

The main outcome of this experiment is that a temporal indication of the presence of aperiodicity representing voiceless excitation leads to significant increments in overall performance in consonant identification. This derives from improved voicing and manner perception. The manner information gained from the signaling of aperiodic excitation can be associated with binary manner features marking affricate, fricative, and plosive consonants, for which voiceless excitation would be expected to be perceptually salient. The amplitude variation of voiceless speech had few effects, but did lead to increased place and plosive information. The effect on place information presumably derives from the use of burst and frication intensity cues.

C. Experiment 2A

This subsidiary experiment was performed for completeness and to allow us to examine a condition representing the timing of both voiced and voiceless excitation in the absence of amplitude variation. This experiment measures the effect of a cue to the presence of voiceless excitation in the absence of amplitude cues, and the effect of amplitude variation when the timing of both voiced and voiceless speech excitation is cued.

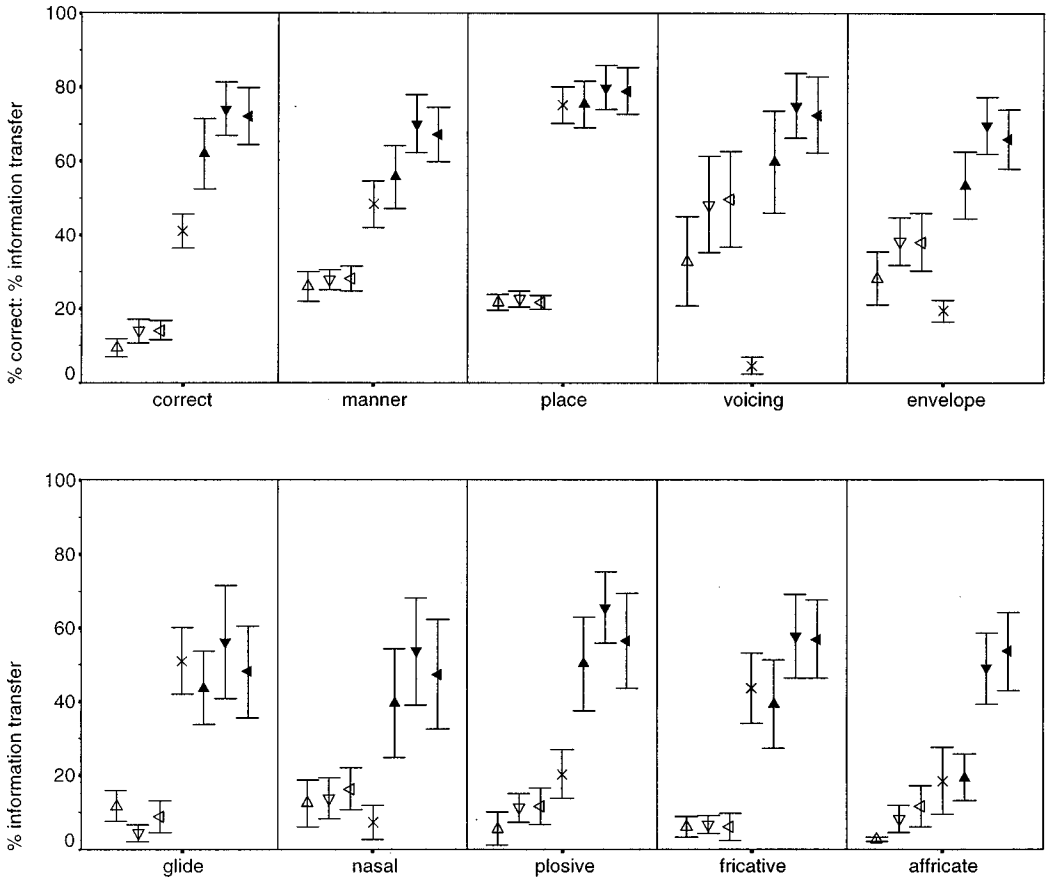

FIG. 4. Upper panel: percent correct and percent information transfer for manner, place, voicing, and envelope features. Lower panel: percent information transfer for binary manner features. Data from experiment 2A, three subjects. Conditions are shown as **F**: △; **F+N**: ▽; **F(A)+N(A)**: ◁; **L**: ×. Audio-visual conditions have the same symbol as the corresponding auditory conditions, but filled. Error bars show ±2 standard errors over 18 data points.

1. Method

Experiment 2A used the same methods as previously, and three naive observers. Seven conditions were used, involving lipreading alone as condition **L**, and three acoustic signals both with and without lipreading: **F**, **F+N**, and **F(A)+N(A)**. Condition **F+N** encoded the presence of voiced or purely voiceless excitation and F_0 with no amplitude variation, and had not been used before.

2. Results and discussion

As before, the last 6 of 10 test lists of 48 consonants were analyzed for each subject. Results for each measure and condition are shown in Fig. 4. One subject showed lower performance overall than the subjects of experiments 1 and 2, and this resulted in a wider range and lower mean scores than previously. The analyses proceeded similarly to that for experiment 2. A series of ANOVAs using factors of condition, session, and subject showed that here neither test session nor its interaction with condition was a significant factor for any measure. The main analyses[4] excluded condition **L** and employed a factorial approach to examine the effects of the presence of visual cues, the presence of a cue to voiceless excitation (comparing signals **F** and **F+N**), and the representation of amplitude envelope [comparing signals **F+N** and **F(A)+N(A)**]. As previously, visual cues led to highly significant increases in overall accuracy and manner, place, voicing, and envelope features. As in experiment 2, the presence of the voiceless excitation cue produced significant effects for all measures except place information; overall accuracy, $F(1,2)=69.13$, $p=0.014$; manner, $F(1,2)=38.19$, $p=0.025$; voicing, $F(1,2)=24.03$, $p=0.039$; envelope, $F(1,2)=137.27$, $p=0.007$. For manner, the effect of the voiceless cue interacted with the presence of visual information $[F(1,2)=141.7, p=0.007]$ and was greater in audio-visual conditions. This interaction term was also significant for place information $[F(1,2)=31.8, p=0.030]$, where the voiceless cue had no detectable effect in auditory conditions, but led to a significant increase in audio-visual scores $[F(1,2)=138.4, p=0.007]$.

There were no main effects of amplitude variation. However, as in experiment 1, there were significant interactions involving amplitude variation and visual information. Here the interactions occurred for overall accuracy $[F(1,2)=334.1, p=0.003]$, manner information $[F(1,2)=205.7, p=0.005]$, voicing information $[F(1,2)=23.91, p=0.039]$, and the envelope feature $[F(1,2)=20.74, p=0.045]$. Owing to the small number of subjects here, ANOVA has insufficient power to detect the small effects of amplitude variation that these interactions suggest may be present here in audio-visual conditions.

As in experiment 2, information transfer for binary affricate, fricative, and plosive manner features (Fig. 4) were significantly affected by the acoustic signal. Both plosive

[$F(1,2)=334$, $p=0.003$] and affricate [$F(1,2)=355$, $p=0.003$] information were significantly increased by the cue to the presence of voiceless excitation. For the affricate feature this effect was, as in experiment 2, much larger in audio-visual conditions. Fricative information also tended to be increased by the voiceless cue, but here only in audio-visual conditions.

D. Discussion

Experiment 2A replicates the finding of experiment 2 that the timing of voiceless excitation can contribute significantly to consonant identity. It shows the same pattern of interaction as previously between visual cues and amplitude variation, whereby audio-visual scores for overall accuracy and manner information are typically higher when amplitude variation was encoded.

II. EXPERIMENT 3: AUDIO-VISUAL SENTENCE PERCEPTION

The third experiment was performed to examine the contribution of these same temporally coded information elements to the audio-visual perception of connected speech. Here, the perception of supra-segmental prosodic information is likely to contribute to performance through lexically and syntactically based stress. While the principal source of prosodic information is generally thought to be intonation, the acoustic correlates of stress also include increased amplitude. Hence it may be expected that amplitude variations are of more importance in connected speech than for consonant identification. Breeuwer and Plomp (1986) have already shown that audio-visual perception of connected speech was significantly improved when amplitude envelope variation was added to a signal that conveyed F_0 information, with syllable correct scores being around 13% higher when amplitude variation was included.

One purpose of this final experiment was to ensure that the signal processing methods used here did not have some artifact that reduced the contribution of speech amplitude variation to consonant identification. In addition, since consonant identification in experiments 2 and 2A was significantly enhanced by the availability of a cue to the presence and duration of voiceless excitation, this final experiment also provides a test of the significance of this factor in sentence-level perception. In the present experiment, the role of F_0 variation was not examined in comparison to a fixed-frequency signal, as the limited number of sentence test lists constrained the number of conditions that could be employed. However, other studies, notably that of Waldstein and Boothroyd (1994), have clearly established that F_0 variation plays a major role in audio-visual sentence perception.

A. Experimental method

1. Conditions

The three auditory supplements employed were **F**, **F(A)**, and **F(A)+N(A)** as above. The fourth condition was lipreading (**L**) with no acoustic signal.

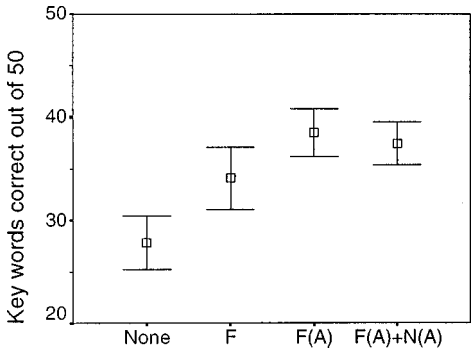

FIG. 5. Mean number of key words identified by condition. The error bars show ±2 standard errors over 30 data points (subject×test session).

2. Speech processing

Speech processing was essentially the same as previously, except that the pulse and carrier signals had a 5 kHz rather than a 400-Hz bandwidth.

3. Speech materials

Speech materials were taken from the UCL EPI audio-visual recording of the BKB sentences (Foster *et al.*, 1993). The speaker was an adult female. Each of the 21 sentence lists comprises 16 sentences, each with 4 or 5 ''key'' content words that are scored for correctness.

4. Subjects

Six normally hearing subjects took part. All were Speech Science students at UCL aged between 21 and 35. All had normal hearing and normal or corrected-to-normal vision. None had taken part in the previous experiments.

5. Procedure

Each subject first received an unscored practice sentence list in each of the four conditions. Because of the limited number of available lists, the same list was used for practice in each condition. This was List 1 which, according to Foster *et al.* (1993), differs most in difficulty from the other 20 lists. Subjects subsequently received five test sessions comprising one test list in each of the four conditions. The order of the four conditions was counterbalanced over five testing sessions. The subjects were spilt into two groups, and two counterbalanced orders were used to distribute lists between the test conditions.

B. Results

Each sentence list was scored according to the ''key-word tight'' (KW-T) procedure (Bamford and Wilson, 1979). This was preferred to the ''key-word loose'' method since it requires the key words to be identified exactly, and hence was expected to reflect more accurately the perception of detailed phonetic information such as the presence of voiceless /s/ in indicating plurality. The group results are shown in Fig. 5.

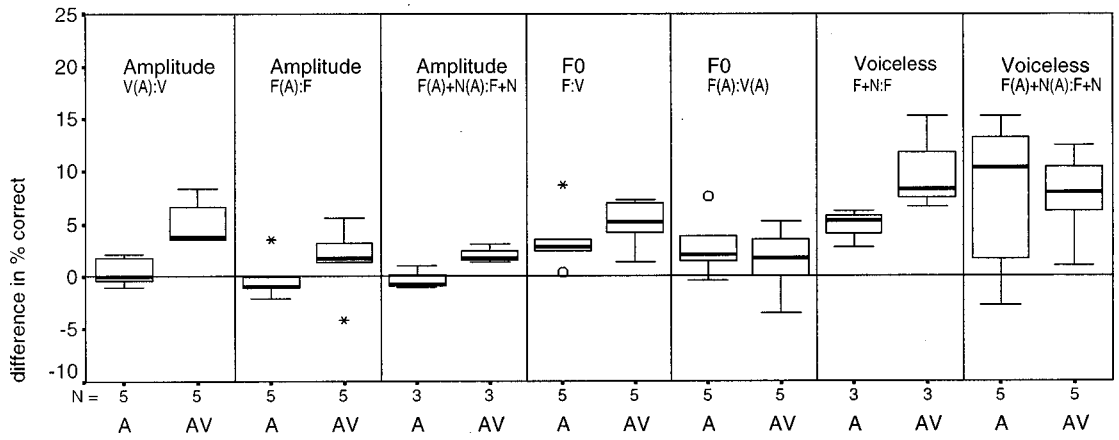

FIG. 6. Differences in % correct scores produced by the addition of amplitude variation (leftmost three panels), F_0 variation (central two panels), and a cue to the presence of voiceless excitation (rightmost two panels). The legends A and AV on the horizontal axis indicate auditory and audio-visual conditions, respectively. The compared auditory signals are indicated in each panel. The plots show the interquartile range (box), the median (bar), and the extreme values (whiskers) of the range of differences in each case. Each data set is from a single experiment. The number of subjects in each data set is also shown (N).

1. Analysis

Because scores from some subjects approached the upper bounds of the test in some conditions, an arcsine transformation was applied to the data prior to a repeated-measures analysis of variance. Comparisons between conditions were made using Tukey HSD tests. As expected, all three of the audio-visual conditions showed significantly higher scores than the visual only condition. Scores in conditions **L+F(A)** and **L+F(A)+N(A)** did not differ significantly, while both showed a significantly higher score than condition **L+F**, the difference being around 7.5 words out of 50. There was also a significant practice effect over sessions $[F(4,20)=10.55, p<0.001]$, but no significant interaction of condition and session.

2. Discussion

In showing a significant contribution of amplitude envelope variation, these results are consistent with others in the literature. Since we have found only a small contribution of amplitude variation at the level of segmental (consonant) perception, we presume the more substantial effect here is partially due to supra-segmental correlates of amplitude variation.

A contribution of voiceless excitation information is not apparent here, despite it being consistently significant in consonant identification. Presumably the contribution it can make in consonantal manner perception and the enhancement of voicing contrasts is made less significant here by the availability of syntactic and lexical context.

III. SUMMARY AND CONCLUSIONS

A. Perceptual contributions of elements of temporal structure

As previous studies also show, auditory alone scores in consonant identification from spectrally invariant signals are always rather low (e.g., Rabinowitz *et al.*, 1992). In audio-visual conditions, however, these auditory signals lead to quite substantial improvements in performance over purely visual presentation. Several clear findings that hold in both audio-visual and auditory consonant identification have emerged. First, the bulk of the useful temporal information comes from the timing of voiced excitation. Second, the timing of voiceless excitation provides an additional significant contribution that provides cues to both voicing and manner of articulation. Smaller contributions can be associated with F_0 variation, which provides a subtle but significant cue to consonant manner, and in audio-visual conditions only, with amplitude variation.

The importance of F_0 information in the perception of connected speech is well-established. Amplitude variation was confirmed to have significance here, as other studies of audio-visual speech perception have already shown (e.g., Breeuwer and Plomp, 1987; Grant *et al.*, 1985, 1991). A similar conclusion was reached for the auditory perception of sine-wave speech by Remez and Rubin (1990). They suggested that amplitude envelope variations were significant as a cue to syllable boundaries, especially around plosive consonants. Owing to the very low performance levels expected from sentence perception in purely auditory conditions with the signals used here, it is not readily possible to make comparable measures of any contribution of amplitude variation in these conditions

The relative effects of temporal cues to aperiodic excitation, F_0, and amplitude variation to consonant identification in experiments 1, 2, and 2A are shown together in Fig. 6. This displays the within-subject differences in performance attributable to the presence or absence of these cues from auditory and audio-visual conditions that differ only in respect of the added information. The median difference in auditory performance produced by the addition of amplitude variation is very close to zero in each comparison. The audio-visual performance increment with the addition of amplitude variation is, however, always greater than zero, but rarely exceeds 5% in any individual subject. F_0 variation led to small performance increments more generally in both auditory and audio-visual presentation, typically of around 3%. Somewhat larger effects ranging up to 15% increases in accuracy were found for the voiceless cue.

That amplitude variation does not, in the absence of

spectral cues, lead to substantial increments in consonant identification may explain the lack of an effect of amplitude compression in consonant identification. For example, Souza and Turner (1996) have shown that the auditory identification of signal-correlated noise stimuli based on consonants is unaffected by compression of the amplitude envelope.

A temporal coding of the timing of voiceless excitation was shown to contribute to consonant identification in both auditory and audio-visual conditions, but not to the audio-visual identification of sentences. This is consistent with Articulation Index importance functions that place greater weight on higher frequency bands for nonsense word materials when phonemes occur with equal frequency than for meaningful connected speech (Studebaker *et al.*, 1987).

B. Temporal cues to consonant features and sentence perception

In terms of perceptual features for consonant identity, we find evidence that voicing information is conveyed by the timing of both periodic and aperiodic excitation. Manner information is also provided by the timing of periodic and aperiodic excitation and here there is in addition a modest but significant contribution from F_0 variation, and from amplitude variation in audio-visual conditions. For binary manner features, the timing of voiceless excitation is a cue for plosive, fricative, and affricate manners, while F_0 and amplitude variations contribute to plosive and nasal information.

The transfer of place of articulation information solely by the auditory signals used here is slight in each case. However, experiment 2 did show a significant increase in place identification due to the amplitude variation of voiceless speech. There is thus no evidence that place-related variations in voice-onset time (Lisker and Abramson, 1967) that would be encoded in the timing of voiced and voiceless excitation lead to salient perceptual cues to place of articulation in the present experiments. However, the amplitude (but not the timing) of voiceless speech does appear to have small effects on place perception.

In the classification of temporal speech features, it seems important to distinguish two components of low-rate *envelope* information, the simple presence of energy, and the perhaps less significant low-rate amplitude variation. Temporal *periodicity* information clearly contributes to consonant perception through contrasting periodic and aperiodic excitation, and also through subtle effects of F_0 on manner perception, in addition to the well known supra-segmental role of F_0.

C. Implications for prosthesis design

Since amplitude variation has only minor effects on the detailed perception of consonantal features, it may be expected that, at least for speech in quiet, considerable compression of amplitude variation could be introduced without substantial cost provided that the typically gross supra-segmental variations in amplitude remain audible. It would therefore be expected that speech receptive prostheses intended for those who largely rely on audio-visual perception need not convey finer details of speech amplitude variation. It is perhaps not surprising in the light of these findings that several studies show amplitude compression to have rather slight effects on speech perception at both segmental and sentence levels (Drullman and Smoorenburg, 1997; Souza and Turner, 1996; van Tasell and Trine, 1996). In noisy environments, however, it is plausible that variations of amplitude may contribute to the ability to segregate speech from noise, and here compression may be more detrimental.

At least where auditory signals lack spectral structure, fundamental frequency information is well-established as a source of useful information in sentence-level audio-visual speech perception. The results of experiment 1 suggest that prostheses which ensure a salient percept of F_0 may also aid in the identification of consonants. That the explicit temporal representation of voiced and voiceless excitation can contribute to consonant identification may also be of practical significance for the design of hearing prostheses, since it would be likely to contribute to the audio-visual perception of low-redundancy messages. Voiceless excited speech is of course distinct from voice excited speech not only by aperiodicity, but also typically by the presence of predominantly higher frequency energy. In listeners for whom higher frequencies are not audible, a temporal coding of periodicity and aperiodicity is likely to be the only possible means of preserving this excitation contrast. Where sufficient frequency range and useful spectral resolution is retained, this contrast may also be accessible from spectral structure, and the significance of temporal cues to aperiodicity when spectral cues are also available merits further investigation.

ACKNOWLEDGMENTS

Supported by TIDE Projects Nos. TP133/206 (STRIDE) and TP1217 (OSCAR), Medical Research Council Grant No. G9020214, and a Wellcome Trust Vacation Scholarship. We are grateful to Kirsti Reeve, Kerensa Smith, and Athena Euthymiades for carrying out the data collection, and to Winifred Strange, James Hillenbrand, and two anonymous reviewers for constructive comments on the manuscript.

[1]The use of the Voiscope for the detection of voiced excitation involves a thresholding procedure on a signal that follows the r.f. impedance across the larynx. This signal gives an unambiguous indication of vocal fold closure, but a less exact indication of smaller quasi-periodic movements of the vocal folds that do not result in vocal fold contact. It is possible for weak laryngeal excitation to be missed. The threshold was set as low as possible without allowing noise from the recorded larynx electrode output to trigger the voicing detector.

[2]The ''pulse'' setting of the VBF8 filter was used. This setting selects a filter type with a maximally flat delay response over frequency. The -3 dB and other cutoff frequencies quoted above are measured values using a nominal 50-Hz cutoff frequency.

[3]Given the small subject numbers, the conventional mixed-effect ANOVA model has been modified to maximize the power of these analyses (experiments 1 and 2). The conventional error term, which is the interaction of the tested factor with subject, was excluded from the ANOVA model when it was clearly nonsignificant ($p>0.2$). Hence the error term for a given test becomes a higher order interaction, which now includes the variability from the excluded interaction. Where each relevant higher order interaction was also clearly nonsignificant, the overall MS error was used. This pooled error has the effect of increasing the error term *df* and the power of each *F* test. Having so modified the ANOVA model, the typical power for a dif-

ference in percent correct scores of around 3%, as for example, the effects in experiment 1 of F_0 variation in auditory and audio-visual conditions, and that of amplitude variation in audio-visual conditions, was between 0.85 and 0.95.

[4]For this experiment, it proved impossible to increase the power of these ANOVAs by discarding nonsignificant interactions with the subject factor, since these interaction terms were generally substantially smaller than the residual error.

Baer, T., and Moore, B. C. J. (**1993**). ''Effects of spectral smearing on the intelligibility of sentences in noise,'' J. Acoust. Soc. Am. **94**, 1229–1241.

Baker, R., and Rosen, S. (**1994**). ''Temporal information in consonant identification,'' Speech Hearing and Language; Work in Progress, Dept. Phonetics and Linguistics, University College London, **7**, 3–29.

Bamford, J., and Wilson, I. (**1979**). ''Methodological considerations and practical aspects of the BKB sentence lists,'' in *Speech-Hearing Tests and the Spoken Language of Hearing-Impaired Children*, edited by J. Bench and J. Bamford (Academic, London), pp. 147–187.

Breeuwer, M., and Plomp, R. (**1986**). ''Speech reading supplemented with auditorily presented speech parameters,'' J. Acoust. Soc. Am. **79**, 481–499.

Drullman, R., and Smoorenburg, G. F. (**1997**). ''Auditory-visual perception of compressed speech by profoundly hearing impaired subjects,'' Audiology **36**, 165–177.

Fant, G. (**1973**). *Speech Sounds and Features* (MIT Press, Cambridge, MA).

Faulkner, A., Rosen, S., and Moore, B. C. J. (**1990**). ''Residual frequency selectivity in the profoundly hearing impaired listener,'' Br. J. Audiol. **24**, 381–392.

Faulkner, A., Ball, V., Rosen, S., Moore, B. C. J., and Fourcin, A. J. (**1992**). ''Speech pattern hearing aids for the profoundly hearing-impaired: Speech perception and auditory abilities,'' J. Acoust. Soc. Am. **91**, 2136–2155.

Foster, J. R., Summerfield, A. Q., Marshall, D. H., Palmer, L., Ball, V., and Rosen, S. (**1993**). ''Lip-reading the BKB sentence lists; corrections for list and practice effects,'' Br. J. Audiol. **27**, 233–246.

Fourcin, A. J. (**1977**). ''English speech patterns with special reference to artificial auditory stimulation,'' in *A Review of Artificial Auditory Stimulation: Medical Research Council Working Group Report*, edited by A. R. Thornton (Institute of Sound and Vibration Research, University of Southampton), pp. 42–44.

Fourcin, A. (**1990**). ''Prospects for Speech Pattern Element Aids,'' Acta Oto-Laryngol. Suppl. **469**, 257–267.

Grant, K. W., Ardell, L. H., Kuhl, P. K., and Sparks, D. W. (**1985**). ''The contribution of fundamental frequency, amplitude envelope and voicing duration cues to speechreading in normal-hearing subjects,'' J. Acoust. Soc. Am. **77**, 671–677.

Grant, K. W., Braida, L. D., and Renn, R. J. (**1991**). ''Single-band amplitude envelope cues as an aid to speechreading,'' Q. J. Exp. Psychol. **43A**, 647–678.

Lisker, L., and Abramson, A. S. (**1970**). ''The voicing dimension: some experiments in comparative phonetics,'' in Proceedings of the Sixth International Congress of Phonetic Sciences, Prague, 1967 (Academia, Prague), pp. 563–567.

Miller, G. A., and Nicely, P. E. (**1955**). ''An analysis of perceptual confusions among some English consonants,'' J. Acoust. Soc. Am. **27**, 338–352.

Moore, B. C. J. (**1996**). ''Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids,'' Ear Hear. **17**, 133–161.

Rabinowitz, W. M., Eddington, D. K., Delhorne, L. A., and Cuneo, P. A. (**1992**). ''Relations among different measures of speech reception in subjects using a cochlear implant,'' J. Acoust. Soc. Am. **92**, 1869–1881.

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

Risberg, A., and Lubker, J. L. (**1978**). ''Prosody and speechreading,'' Report of STL-QPSR, Dept. of Linguistics, University of Stockholm, Stockholm, Sweden, **4**, 1–16.

Rosen, S., Fourcin, A. J., and Moore, B. C. J. (**1981**). ''Voice pitch as an aid to lipreading,'' Nature (London) **291**, 150–152.

Rosen, S., Faulkner, A., and Smith, D. A. J. (**1990**). ''The psychoacoustics of profound hearing impairment,'' Acta Oto-Laryngol. Suppl. **469**, 16–22.

Rosen, S. (**1992**). ''Temporal information in speech: Acoustic, auditory and linguistic aspects,'' Philos. Trans. R. Soc. London, Ser. B 367–373.

Shannon, R. V. (**1993**). ''Psychophysics,'' in *Cochlear Implants: Audiological Foundations*, edited by R. S. Tyler (Singular Publishing, San Diego), pp. 357–388.

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

Shinn, P., and Blumstein, S. E. (**1984**). ''On the role of amplitude envelope for the perception of [b] and [w],'' J. Acoust. Soc. Am. **75**, 1243–1252.

Souza, P. E., and Turner, C. W. (**1996**). ''Effect of single-channel compression on temporal speech information,'' J. Speech Hear. Res. **39**, 901–911.

Studebaker, G. A., Pavlovic, C. V., and Sherbecoe, R. L. (**1987**). ''A frequency importance function for continuous discourse,'' J. Acoust. Soc. Am. **81**, 1130–1138.

Summerfield, A. Q. (**1987**). ''Some preliminaries to a comprehensive account of audio-visual speech perception,'' in *Hearing by Eye: The Psychology of Lipreading*, edited by B. Dodd and R. Campbell (Lawrence Earlbaum Associates, Hove, U.K.), pp. 3–51.

Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. (**1987**). ''Speech waveform envelope cues for consonant recognition,'' J. Acoust. Soc. Am. **82**, 1152–1181.

Van Tasell, D. J., Greenfield, D. G., Logemann, J. J., and Nelson, D. A. (**1992**). ''Temporal cues for consonant recognition: Training, talker generalization, and use in the evaluation of cochlear implants,'' J. Acoust. Soc. Am. **92**, 1247–1257.

Van Tasell, D. J., and Trine, T. D. (**1996**). ''Effects of single-band syllabic amplitude compression on temporal speech information in nonsense syllables and in sentences,'' J. Speech Hear. Res. **39**, 912–922.

Waldstein, R. S., and Boothroyd, A. (**1994**). ''Speechreading enhancement using a sinusoidal substitute for voice fundamental frequency,'' Speech Commun. **14**, 303–312.