VOICING, FUNDAMENTAL FREQUENCY, AMPLITUDE ENVELOPE AND VOICELESS-NESS AS CUES TO CONSONANT IDENTITY

Stuart Rosen, Andrew Faulkner, Kirsti Reeve* and Kerensa Smith**
Northwestern University, Evanston, Illinois, U.S.A.
**Department of Phonetics and Linguistics, University College London, U.K.*

ABSTRACT

Of long-standing theoretical and practical interest is the extent to which cues to consonant identity can be provided by purely temporal auditory features (periodic and aperiodic excitation, and amplitude envelope). Here we show that the primary features used by normal observers involve the on-and-off patterning of silence, periodicity and aperiodicity (both with and without lipreading). Additional variations in envelope and fundamental frequency provide little further information.

INTRODUCTION

For some years, we have been developing a speech-pattern hearing aid for profoundly hearing-impaired people [1]. The original SiVo aid (**Si**nusoidal **Vo**ice) extracted from the speech only the voice fundamental frequency (Fx) and presented it as a sinusoid at a constant loudness — a signal which provides an effective auditory supplement to lipreading. Now, further speech pattern elements are being incorporated, representing the speech amplitude envelope and voiceless frication, and we wanted to obtain normative data for comparison with results from our hearing-impaired listeners. Not the least reason for this is practical — to know if our current tests would be sensitive to these extra acoustic features.

Having decided to focus on the perception of intervocalic consonants for the moment, there were other interesting issues to address. For example, it has often been noted that Fx contours have a microstructure that could, in theory, transmit segmental information over and above that contained in the simple on-and-off pattern of voicing.

There is also currently much interest in the temporal structure of speech [2, 3] and in particular, the degree to which amplitude envelope is important. At least one source of this interest is the extent to which amplitude compression in auditory prostheses, with its transformation of the natural envelope of speech, would have a deleterious effect. Here we compare the most extreme compression (signals with no variation in amplitude when "on") to signals with natural variations in envelope.

EXPERIMENT 1

Experiment 1 investigated a simple coding of the voiced components of speech only. Represented were the on-off pattern of larynx excitation, its' fundamental frequency, and the amplitude envelope. The key questions were whether Fx variation and/or envelope provide cues to consonant identity beyond those in the on-off pattern of larynx excitation.

Methods

A total of 9 conditions were tested: lipreading alone (**L**), plus 4 sound conditions with (**L+**) and without lipreading:

V - A fixed-frequency, fixed-amplitude signal indicating vocal fold vibration.

V(A) - as for **V** but with added amplitude envelope, derived from the original speech.

Fx - A fixed-amplitude signal whose periodicity followed the speaker's Fx.

Fx(A) - as for **Fx**, but with an amplitude envelope added.

Four normal-hearing native speakers of British English took part. Speech materials comprised each of the 24 English consonants between the vowel /ɑ/. Five distinct video-recorded lists (female speaker) were employed, each consisting of 2 tokens of each consonant. One list was reserved for initial training in each condition, whilst 4 were used for testing.

Fundamental frequency and voicing information were recorded by means of an electro-laryngograph on the speaker's throat at the time of recording, in the form of narrow pulses synchronised to the speaker's vocal fold closures. These pulses were then used as input to an external device with two modes of operation for generating the test sounds. For conditions involving **Fx**, the original pulses were used to trigger other pulses on a 1-for-1 basis. For conditions involving **V**, the original pulses were used to gate on and off a train of pulses of constant frequency. The triggered pulses were low-pass filtered at 400 Hz (18 dB/octave) to make them pleasant to listen to. For conditions with amplitude information, envelopes were derived by full-wave rectifying the 3-kHz low-pass filtered speech, and smoothing the result with a 30 Hz low-pass filter. These were then multiplied against the appropriate pulse train. All signals were recorded for testing purposes, and presented free-field using a loudspeaker.

Analysis

Each session was analyzed separately by constructing a confusion matrix from which overall proportion correct scores were derived, together with unconditional information transfer measures for:

voicing: voiced vs. voiceless

place: bilabial vs. labiodental vs. dental vs. alveolar vs. palatal vs. velar vs. pharyngeal

manner: plosive vs. affricate vs. fricative vs. nasal vs. glide

voice/man: a slightly collapsed voicing/manner feature, closely related to so-called envelope features [3] — voiced plosives vs. voiceless plosives vs. voiceless fricatives vs. sonorants (nasals + glides) vs. voiced fricatives.

To allow for learning, only the last 6 sessions for each condition of the 10 run were analyzed. Statistical claims are made on the basis of an ANOVA including an observer x condition interaction (which was often significant), and Tukey's Studentized Range Test ($p \leq 0.05$).

Results

Table 1 shows mean performance as a function of condition. Values with a common symbol in the same column (*, #, @) are indistinguishable statistically. Although more information tends to lead to better performance, neither fundamental frequency nor envelope increase performance very much compared to on-off voicing. That Fx variations aid consonant identification little has already been shown [4], but the small effects of envelope variation come as a surprise.

TABLE 1

condition	*feature*				
	correct	*voice/man*	*voicing*	*manner*	*place*
V	# 13	# @ 45	@ 68	# 28	# 22
V(A)	# 14	# @ 48	@ 69	# 30	# 23
Fx	# 18	@ 52	@ 75	# 35	# 25
Fx(A)	# 17	# @ 50	@ 72	# 33	# 24
L	54	# 43	15	60	* 90
L + V	* 79	* 71	* 92	@ 67	* 93
L + V(A)	* 83	* 76	* 93	@ * 73	* 95
L + Fx	* 83	* 77	* 94	* 75	* 93
L + Fx(A)	* 85	* 78	* 95	* 77	* 94

EXPERIMENT 2

Experiment 2 was primarily concerned with the role of voiceless frication and envelope. There were 3 different sound signals, presented with and without lipreading, making for a total of 6 conditions. Apart from **Fx(A)** used in Experiment 1, the other sound signals were:

Fx(A)+Nz - as for **Fx(A)** above, with a band of fixed-level noise present during periods of voiceless excitation.

Fx(A)+Nz(A) - as above, but with an amplitude envelope on the noise as well.

Methods

Five new observers took part, following the same procedure as in Experiment 1. Signal components reflecting voicing in the speech signal were created as described for Experiment 1. Voiceless excitation was detected by a spectral balance circuit comparing the amount of energy above and below 3 kHz in the speech signal. However, the presence of voice pulses from the laryngograph overrode the comparator. Thus, voiceless excitation could only be detected in the absence of voicing. When the comparator indicated voiceless excitation, it gated a white noise that was then mixed with the voicing pulses, for final low-pass filtering at 400 Hz. For conditions with amplitude information, envelopes were derived by full-wave rectifying the broad-band speech signal and smoothing the result using a 30 Hz low-pass filter. These envelopes were then multiplied against the white noise. Both the noise and pulse train signals were low-pass filtered at 400 Hz. Again, all signals were recorded for testing purposes.

Results

Analysis procedures were the same as described for Experiment 1, resulting in the summary found in Table 2. Again, more information tends to lead to better performance. The addition of voiceless information almost always leads to significantly improved performance (except for the place feature). However, the addition of amplitude envelope never causes significant increments for the features analysed (just as found in Experiment 1).

TABLE 2 condition	*feature*				
	correct	*voice/man*	*voicing*	*manner*	*place*
Fx(A)	# 19	# 47	62	# 32	# 24
Fx(A) + Nz	@ 24	@ 60	# 72	@ 45	# 24
Fx(A) + Nz(A)	@ 27	@ 61	@ # 79	@ 45	# 27
L + Fx(A)	68	68	@ 82	61	* 79
L + Fx(A) +Nz	* 76	* 82	* 92	* 73	* 80
L + Fx(A) +Nz(A)	* 75	* 79	@ * 88	* 72	* 82

EXPERIMENT 3

Experiment 3 focused primarily on the overall role of envelope, using the same methods as previously, but with three new observers. Seven conditions were used, involving lipreading alone, plus three sound signals both with and without lipreading: **Fx**, **Fx+Nz** (which had not been used previously), and **Fx(A)+Nz(A)**.

Results

The results (Table 3), lead to essentially the same conclusions as the previous two experiments. Variations in envelope beyond a simple binary indication of amplitude never lead to statistically significant increments in performance. But the addition of voiceless information often does, especially for voicing and for other features in conjunction with lipreading.

TABLE 3 condition	*feature*				
	correct	*voice/man*	*voicing*	*manner*	*place*
Fx	# 9	# 33	# 33	# 26	# 21
Fx + Nz	# 14	# 39	@ 50	# 28	# 22
Fx(A) + Nz(A)	# 14	# 39	@ 48	# 28	# 22
L	@ 41	# 33	5	@ 48	@ 77
L + Fx	62	58	@ 60	56	@ * 79
L + Fx +Nz	* 72	* 69	* 73	* 67	@ * 81
L + Fx(A) +Nz(A)	* 74	* 73	* 75	* 70	* 82

DISCUSSION

At first sight, these results bode well for auditory prostheses that distort envelope. Insofar as envelope variations made little difference to performance, it is clear that the bulk of temporal segmental information is contained in the on-and-off patterning of silence, periodicity and aperiodicity.

But there are two important caveats. First, it is not possible to extrapolate to connected discourse from results with consonants. We already know that variations in Fx aid consonant identification little in comparison to a simple voicing indicator, even though such variations are of great utility in connected discourse [5]. And there is evidence that envelope variations are as much a benefit in connected discourse as are Fx variations [6]. Second, it may not be wise to extrapolate from the results of normal observers to impaired ones. Faulkner *et al.* [1] have already shown that for some profoundly hearing-impaired observers listening to signals analogous to those used above, the addition of envelope can be of significant benefit, both in connected discourse *and* in consonant identification. It may be that impaired listeners are less able than our normally hearing listeners to use the on-off timing of voiced and voiceless excitations, and in consequence, may depend more on the use of other correlated cues conveyed by amplitude envelope.

ACKNOWLEDGEMENTS

Supported primarily by the Medical Research Council (UK), with important additional support from a Wellcome Trust Vacation Scholarship, CEC TIDE projects 206 (STRIDE) and 1217 (OSCAR), and Northwestern University.

REFERENCES

[1] Faulkner A *et al.* (1992), "Speech pattern hearing aids for the profoundly hearing-impaired: Speech perception and auditory abilities", *J Acoust Soc Am*, vol. 91, pp 2136-2155.

[2] Rosen S (1992), "Temporal information in speech: acoustic, auditory and linguistic aspects", *Phil Trans Royal Soc London B*, vol. 336, pp 367-373.

[3] Van Tasell DJ *et al.* (1992), "Temporal cues for consonant recognition: Training, talker generalization, and use in the evaluation of cochlear implants", *J Acoust Soc Am*, vol. 92, pp 1247-1257.

[4] Rosen S *et al.* (1979), "Lipreading with fundamental frequency information", *Proc Inst Acoust Autumn Conf*, pp 5-8.

[5] Rosen S *et al.* (1980), "Lipreading connected discourse with fundamental frequency information", *Brit Soc Audiol Newsletter (Summer)*, pp 42-43.

[6] Grant KW *et al.* (1985), "The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects", *J Acoust Soc Am*, vol. 77, pp 671-677.