CUES TO THE RECOGNITION OF SOME LINGUISTIC FEATURES OF WHISPERED SPEECH IN ENGLISH

J. TRIM

The study of whisper, as of other abnormal speech mechanisms, is of interest for the underestanding of the relation of speech to language. If it is considered that speech perception is a matter of an automatic response to certain acoustic constants, whether absolute or relational, in the speech wave, necessarily produced by automatized articulations, phonological units will be defined in strictly phonetic terms. In the alternative view, it may be held that, where speaker and listener share the same linguistic competence, it is the task of the speaker under abnormal conditions to incorporate into the speech-wave he produces in a particular performance sufficient information for the listener to identify the text intended; and it is the listener's task to make that identification on the basis of the clues available.

In whisper, vocal cord vibration is avoided, the periodic voice source being replaced by an aperiodic voice source. The energy supplied over a wide frequency band is filtered in accordance with the resonant characteristics of the vocal tract, as in voice. The spectral features characteristic of many speech-sounds are thus unchanged. The first and second vowel formants, for instance, are as well-marked and perhaps more stable for not being subject to fluctuations in harmonic intervals and the same applies to consonants primarily characterized by formants and formant transitions. The local supraglottal noise effects characterizing plosives; affricates and fricatives are unaffected, but the question of the voiced-voiceless distinction arises. In most cases, concomitant features, well-known from phonetic descriptions, take over.

In final position the cues are: length of preceding vowel, lateral or nasal; in the case of plosives: length of stop, length and intensity of the noice burst or release; in the case of fricatives; the length and intensity of friction. Intervocalically in disyllables the length of the stop or friction is dominant. If no substantial degree of lengthening is present an intended "voiceless" consonant is likely to be heard as "voiced" e.g. "latter" as "ladder". Initially the cues are again length and intensity, and aspiration. In whisper aspiration cannot of course be equated with delayed onset of voice. Instead, the onset of a vowel-like segment with strong concentration of energy in the formant is separated from the initial consonant by a segment with diffuse noise centred on 4 kc. The presence of this diffuse noise, as opposed to the abrupt onset of whisper with clear formant structure (especially initiated by a glottal plosive) characterizes h-words as opposed to V-words. There do not appear to be cases in

which the use of whisper for voice leads to lexical homophony. The residual features employed are often accentuated in whispered speech, but still the demands on the listener are increased, with a resultant lowering of recognition scores.

By comparison with such lexical features, the intonation of whispered speech presents particularly interesting problems. It has been asserted that the concept of intonation is inapplicable to whisper, since the vocal cords do not vibrate, and the speech wave, being aperiodic, has no fundamental frequency. It is said that in that case the voice has no pitch, and that tone and intonation are excluded. Returning, however, to the first point made in this paper, we may reply that if intonation is part of the meaningful linguistic form of the utterance, speaker and listener will readily adapt to an alternative manifestation.

To test this, a small experiment was conducted along the following lines. Two lists of sixty words each were constructed, in which each member of one of two sets of ten monosyllables, to some extent phonetically balanced, was combined with each of 6 nuclear tones (low fall, high fall, low rise, high rise, fall-rise and rise-fall) in an arbitrary order. The lists were recorded by the experimenter in the order: set I, voiced (test A); set II, whispered (test B); set II, voiced (test C); set I, whispered (test D). The sequence of tones was arbitrarily different in each list. A small panel of trained, professional phoneticians were then asked to write down immediately against the printed word the tone they thought to have heard. The words were presented on tape, at a rate which would allow just enough time for the subject to note the tone, but not to reflect. One or two subjects complained of the speed and may conceivably have gained higher scores if allowed longer. After a short rest, the subjects repeated the test, to see whether any short-term learning had occurred. Two of the subjects then repeated the whole test after six weeks, to see whether the learning effect, if any, had persisted.

The results of the experiment are summarized in the following tables:

I. Tones correctly recognized:

Subject	*A*1	*B*1	*C*1	*D*1	*A*2	*B*2	*C*2	*D*2	Total (480)
A	45	30	49	22	52	29	49	29	305
B	49	33	51	27	53	29	53	33	328
C	50	26	52	24	52	36	58	34	332
D1	53	31	57	30	58	40	57	30	356
E	58	32	59	31	58	44	60	38	380
F1	58	39	60	38	57	46	60	45	403
G	59	46	60	45	59	49	59	54	431
D2	57	36	58	39	57	44	60	40	391
F2	56	50	60	44	56	50	60	44	420
Total	485	323	506	300	502	367	516	347	

II. Breakdown of recognition scores by word and tone:

	no	there	now	oh	yes	six	would	fine	loose	out	Total (180)	
\1	5	11	17	18	16	11	12	11	15	10	126	346
\2	14	17	15	13	15	15	4	16	14	13	136	
/3	15	6	14	12	18	9	18	13	4	6	115	
/4	3	3	4	2	10	8	6	1	11	1	49	
\/5	12	15	14	15	14	17	13	9	12	13	134	
/\6	17	13	12	9	14	15	14	9	16	11	130	
Total (108)	66	65	76	69	87	75	67	59	72	54	690 (774)	

b) Test d

	so	nine	you	nice	will	Franks	oh	where	shut	good	Total	
1	17	12	18	11	15	14	14	16	13	7	137	349
2	7	14	6	13	7	17	14	8	16	7	109	
3	1	10	15	17	11	5	14	18	12	13	116	
4	5	5	4	4	5	2	4	13	2	4	48	
5	14	8	13	8	17	6	17	13	13	14	123	
6	15	16	14	13	8	9	14	6	9	12	116	
Total	59	65	70	66	63	53	77	74	65	57	649 (752)	

c) Scores for attempt 1: 1614 (max 2160)
attempt 2: 1732

d) Scores for test A. 987 (max 1080)
B. 690
C. 1022
D. 647

e) Scores for Voice 1 991 (max 1080)
Whisper 1 623
Voice 2 1018
Whisper 2 714
Voice 1 + 2 2009 93% (max 2160)
Whisper 1 + 2 1337 62% of max, 66% of voiced
Whisper 1 + 2 given interchangeability of high and low fall 1526 75%.

f) Long Term effect: subjects D and F, take 1 759 (max 960)
take 2 811

Though the number of subjects involved in the experiment is small, certain trends seem clear and relatively constant. Small but constant differences of perfor-

mance between subjects persist throughout the test, reflecting the relative skill and experience of the subjects in this field. A small but steady improvement of performance occurs in the voiced tests, flattening off as the 100 % level is approached or reached. The recognition scores for whisper are lower, but far too high, in all cases, to result from chance. The improvement between the first and second attempts is steeper for whisper. There was, however, a slight drop between tests B and D, on both test and repeat. A fatigue effect is perhaps responsible. The figures for the retake, 6 weeks later, by subjects D and F show that the learning effect had persisted.

It should be emphasized that the subjects received no feedback concerning their performance, so that learning proceeded only by the subject sensitizing himself to the recognition cues present.

Some error patterns in whisper recognition were very consistent. The nature of the lexical material had a restricted effect, more attributable to the frequency of use of words in monosyllabic utterances than to their phonetic consistency. Low and high falls were rarely falsely identified except as each other. Indeed, if in view of the gradient nature of falling intonations we allowed the interchange of whispered falls, their recognition rate would rise to 96 % for test B and 97 % for test D. A particularly striking feature of the results is that the high rise, with recognition scores of 96 % in test A and 97 % in test B, has a score in whisper of only just over 25 %, being mainly identified as high fall, low rise or rise fall.

Spectrograms were then made of the 240 items on the test, and examined in an attempt a) to identify the cues to the recognition of tones, b) to relate the distribution of judgements for particular items to the variation in the acoustic properties of the utterance, c) to determine whether the cues were ones present in voiced speech, or new features specific to whisper.

There is no space to deploy the evidence in detail here. The apparent answers to the questions are that Meyer-Eppler's observations, i.e. that the amplitude contour has the shape of the perceived pitch curve, and that the frequency of the third formant rises and falls in conformity with perceived pitch are largely confirmed, though 3rd formant bending is often minimal, and disregarded by listeners when it is in conflict with amplitude contour. This latter feature exists to some extent in voiced speech, as a modification of a general steady diminution of energy in the vowel segment. In whisper, the features are heightened, though still not entirely cancelling out the diminution. It is this feature listeners appear to rely on initially, which accounts for the much higher recognition rate of falls than rises, and the identification of the high rise as high fall or rise fall.

In addition to the features noted, use is made of further spectral features. In some cases, the vowel formants are considerably affected, as in "you". More often, the region up to 2·5 kc is relatively unaffected, and pitch effects are produced by fluctuation in the amount of energy contained in formants 3, 4 and 5, between 2·5 and 4—occasionally—5 kc. The presence of amounts of energy in that area, usually concentrated in formant bands, but sometimes more diffuse, occurs in high pitched

whisper and its absence in low pitched whisper. Falls, rises, fall rises and rise falls are cued by successions of these states. Low fall and high fall, low rise and high rise are differentiated by the relative intensity of this high frequency noise. It is to these features, absent from the spectrograms of the voiced items in the tests, that listeners are learning to respond, rapidly and, it seems, with persistent effect, as their recognition scores improve.