SOME EFFECTS ON INTELLIGIBILITY OF INAPPROPRIATE TEMPORAL RELATIONS WITHIN SPEECH UNITS

A. W. F. Huggins, Bolt Beranek and Newman Inc, 50 Moulton Street, Cambridge, Mass 02138, U. S. A.

The purpose of this paper is to make two arguments. The first is that, despite several failures to find such effects, badly disturbed speech timing, such as occurs often in the speech of the deaf for instance, is a sufficient cause for catastrophic loss of intelligibility. If the timing is sufficiently disturbed that the listener cannot identify the pattern of stressed syllables in the sentence -- or, perhaps, its rhythmic pattern -- the sentence will be unintelligible even though virtually all of the phonemes are clearly identifiable in subsequent listening. If the listener perceives a stress/rhythmic pattern that is different from that intended by the speaker, he is "garden-pathed" away from the correct utterance, and is not able to recode the individual phonemes into the words they represent before they fade from auditory short-term memory.

The second argument is that a reason for earlier failures to find strong relationships between timing and intelligibility is that a listener cannot estimate the effect of a particular timing distortion on speech intelligibility if he knows what the sentence says. This fact is already well known. It forms the basis of a popular way of impressing an audience with the fidelity of a speech vocoding system: a demonstration tape is prepared in such a way that the audience already knows what the test sentence is before they hear it as processed by the system whose performance is to be proved. What is not so well known is how easy it is to fall into the trap set by this fact. To be blunt, although I was very aware of the effect, I fell into the trap (Huggins, 1978), and if it can happen to me, it can happen to anyone!

Speech of the Deaf

A major reason for trying to understand speech timing is the need to improve the intelligibility of deaf speakers. Faulty timing has been implicated in poor intelligibility by virtually every major study of deaf speech this century, but this knowledge has not led to the development of effective training methods.

The most frequently cited ways in which the timing of deaf speech differs from normal speech are (1) slower overall rate; (2) more and longer pauses, often inappropriately placed; (3) inadequate differentiation of stressed and unstressed syllables; and (4) excessive lengthening of some segments, especially stops and fricatives (e.g. Nickerson, 1975). Let us consider the foregoing factors in order. Deaf speakers normally take much longer to produce a specified utterance than do normal-hearing speakers. But to the extent that the slower rate is a result of linear stretching of the time scale, slower speech should be more rather than less intelligible. One usually speaks slower (and also more precisely) to someone who has difficulty understanding, such as a child or a foreigner. Furthermore, when recorded speech is instrumentally expanded in time by a factor of four, intelligibility is not affected although the speech becomes tedious to listen to.

Similarly, it would be very surprising if the addition of appropriately placed pauses had a degrading effect on intelligibility. Pauses can be used to mark explicitly the boundaries between groups of syntactically related words. Boundaries so marked need not be inferred from more subtle cues, and the presence of syntactically appropriate pauses should therefore simplify rather than complicate reception. Further, the pauses effectively give the listener additional time to decode the message, and this too lightens rather than increases the processing load (Aaronson et al, 1971).

The occurrence of inappropriate pauses raises a different issue. Inappropriate pauses occur also in normal speech, where they are interpreted as hesitation pauses. These do not appear to interfere with intelligibility. However, listeners are much more sensitive to the presence of inappropriate than appropriate pauses, the threshold for their detection being almost five times smaller (Boomer and Dittmann, 1962). Presumably, then, if inappropriate pauses were interpreted as hesitation pauses in deaf speech also, no damage would result. Problems would arise, however, if the inappropriate pauses were interpreted as appropriate pauses, because this would signal incorrect segmentation of the message. This argument leads to rather a

different view of how timing errors might interfere with intelligibility: they might introduce misleading information about the message which, once accepted, could not be discarded.

There are other aspects of deaf speech which support such a view. Due to difficulties in coordinating different articulators, deaf speakers often produce sounds extraneous to the required sequence, particularly in making and releasing stops and fricatives (Hudgins and Numbers, 1942). If the listener accepts these extraneous sounds as segments, he cannot then go back and delete them. The perceptual apparatus is very good at filling in missing information, but it is very bad at discarding extraneous information unless it occurs as part of a separate auditory "stream" (Bregman and Campbell, 1971). Thus, listeners will swear that they heard a particular segment in a sentence even though it had been totally removed and replaced with an extraneous sound such as a cough (Warren et al, 1969). But the cough cannot be located in the sentence with any accuracy, since it cannot be integrated into a single stream with the speech. When wanted and unwanted segments arrive in a single auditory stream, as they often do in deaf speech, the listener cannot selectively accept the wanted and reject the extraneous segments, even if he had some way of so classifying the segments as they arrived. van Noorden (1975) has shown that two melodies in the same pitch range cannot be identified if they are played by interleaving the notes from the two melodies. The listener cannot decide to listen to alternate notes. On the contrary, he hears only a single sequence. But if one melody is gradually raised in pitch, the two melodies eventually split into two streams, permitting one to be ignored so that the other melody can be recognized.

The listener is not able to discard some of the information after it has been processed, either, and recent models of speech perception offer an explanation. Jarvella (1971) has shown that the accuracy of a listener's verbatim memory for a continuously presented message shows a sharp drop at the preceding clause boundary, as if the need to keep the raw acoustic data available in short-term memory ends when the clausal material is successfully parsed. Thus, any misinterpretations of the

preceding clause that become apparent later cannot easily be corrected, since the verbatim material necessary to the correction has been deleted from short term memory. Furthermore, if the received sequence of segments fails to trigger recognition of a word, the segments fade quite rapidly from auditory short term memory.

When the foregoing arguments are put together with the known importance of correct stress patterns for recognition of words, the poor intelligibility of deaf speech becomes much easier to understand. The pattern of stresses in a word or phrase is of critical importance to its correct recognition. In fact, there is evidence that listeners will discard correctly-heard segmental cues which they cannot reconcile with the perceived stress pattern. English listeners trying to identify English words and phrases, spoken with inappropriate stress patterns by Indian speakers, consistently produced words that matched the incorrect stress patterns, while correct phonemes occurred in enough of the responses to demonstrate that the necessary segmental cues were in fact present (Bansal, 1966). Second, it is known that timing is a vital cue in the perception of stress, outweighing both intensity (loudness) and pitch (Fry, 1958).

Yet it is not clear how much deaf speakers know about stress patterns. For normal listeners, the stress pattern of a word is centrally involved in its memory coding (Brown and McNeill, 1966). It is unlikely that the deaf use a similar coding without being explicitly taught it. Deaf children do not code letters, presented visually in an immediate recall task, in terms of their auditory and articulatory properties, as do normal hearing children and adults (Conrad and Rush, 1965). If the deaf subjects do not use an auditory or articulatory coding scheme for segments, it is very likely that they also use a different coding scheme for stress patterns -- if, indeed, they have a coding scheme for stress patterns at all. Unless the stress pattern of a word is a central part of its representation in memory, the stress pattern is not likely to be reflected in the required pattern of syllable timing when the word is spoken. Yet this pattern of syllable timing is crucial to the intelligibility of the word for hearing listeners.

There are two aspects of incorrect timing that should be distinguished. One type can be traced directly to the difficulty of programming a rapid sequence of articulations. Timing errors become more frequent and more severe as the sentence to be uttered is made more difficult to articulate. The remedy may lie in trying to teach words as integrated motor patterns, and practicing their production first in isolation and then by substituting them in overlearned phrase or sentence frames. This is particularly important in the case of function words, whose fluency in deaf speech is a major determinant of intelligibility (Monson and Leiter, 1975). Timing errors of the foregoing type could be labeled errors of performance, since the deaf speaker is presumably at least partly aware that his production has fallen short of what was intended. The other aspect of incorrect timing is more important, and errors of this type could be labeled errors of intention. Errors of intention occur if the deaf speaker's model of how speech should be timed is different from that of a hearing speaker. In particular, the model may not incorporate the rules for assigning relative stress levels, and for realizing these in timing patterns.

Some evidence supporting the importance for intelligibility of differentiating stressed and unstressed syllables has been reported by Osberger (1978). She produced slight improvements in intelligibility by editing deaf speech waveforms to correct inadequate differentiation of stressed and unstressed syllables. Her method, however, was unable to separate errors of performance from errors of intention, which may account for the smallness of her effects. Also, she reported no attempt to relate the magnitude of the timing corrections made in individual words to the resulting changes in intelligibility.

I have reported elsewhere a preliminary attempt to measure the effects of errors of intention uncontaminated by errors of performance, using synthetic speech (Huggins, 1978). Simple sentences were synthesized in two versions. In one, stress was correctly assigned, and in the other, unstressed syllables were assigned primary stress, and vice versa. Syllables with secondary stress were not affected. Since the same set of synthesis rules were used for stressed as for unstressed

syllables, any errors of performance that were inherent in the synthesis procedure should have affected the normal and mis-stressed versions equally. But when stress was wrongly assigned, word intelligibility fell from 85% to 50%, and the percentage of sentences "substantially understood" fell from 75% to 25%. The results were not uniform across test sentences, in part because the sentences differed in the proportion of syllables carrying primary, secondary, and un- stress, and in part because of some residual errors in phonetic transcription of the test sentences (which may well account for the less than perfect intelligibility of the normally stressed versions). I hope to correct some of these weaknesses in time for the meeting.

Finally, I want to repeat an anecdote from the study. I have tried several times to make a tape demonstrating how unintelligible speech can become when its timing is wrong, but I have never been satisfied with the results. In fact, I began to wonder if what I was trying to show was true. But when I played the latest tape to a colleague, looking for sympathy, he found it totally unintelligible. The difference between us was that I knew what each test sentence said, and therefore knew its stress pattern, whereas he did not. I would never have run the formal experiment but for his unexpected reaction. How many interesting timing effects have been overlooked, or regarded as too slight to be of interest, for similar reasons?

References

Aaronson, D., N. Markowitz, and H. Shapiro (1971): "Perception and immediate recall of normal and "compressed" auditory sequences," Perception and Psychophysics, 9, 338-344.

Bansal, R. K. (1966): The intelligibility of Indian English: measurements of the intelligibility of connected speech, and sentence and word material, presented to listeners of different nationalities, Unpublished Ph. D. Thesis, London University.

Boomer, D. S. and A. T. Dittmann (1962): "Hesitation pauses and juncture pauses in speech," Language and Speech, 5, 215-220.

Bregman, A. S. and J. Campbell (1971): "Primary auditory stream segregation and perception of order in rapid sequences of tones," J. Experimental Psychology, 89, 244-249.

Brown, R. and D. McNeill (1966): "The tip of the tongue phenomenon," J. Verbal Learning and Verbal Behavior, 5, 325-337.

Conrad, R. and M. L. Rush (1965): "Nature of short-term memory encoding by the deaf," J. Speech and Hearing Disorders, 30, 335-343.

Fry, D. B. (1958): "Experiments on the perception of stress," Language and Speech, 1, 126-152.

Hudgins, C. V. and F. C. Numbers (1942): "An investigation of intelligibility of speech of the deaf," General Psychology Monograph, 25, 289-392.

Huggins, A. W. F. (1978): "Speech timing and intelligibility," in J. Requin (ed.), Attention and Performance VII, Hillsdale, N.J.: Erlbaum.

Jarvella, R. (1971): "Syntactic processing of connected speech," J. Verbal Learning and Verbal Behavior, 10, 409-416.

Monson, R. B. and E. Leiter (1975): "Comparison of intelligibility with duration and pitch control in the speech of deaf children," J. Acoust. Soc. Amer., 57, S69 (A).

Nickerson, R. S. (1975): "Characteristics of the speech of deaf persons," Volta Review, 77, 342-362.

Osberger, M. J. (1978): The effect of timing errors on the intelligibility of deaf children's speech. Unpublished doctoral thesis, City University of New York.

van Noorden, L. P. A. S. (1975): Temporal coherence in the perception of tone sequences. Eindhoven, Netherlands: Technische Hogeschool (Doctoral thesis).

Warren, R. M., C. J. Obusek, R. M. Farmer, and R. P. Warren (1969): "Auditory sequence: confusions of patterns other than speech or music," Science, 164, 586-587.