PERCEPTUAL INTEGRATION OF RISE TIME AND SILENCE IN AFFRICATE/FRICATIVE AND PLUCK/BOW CONTINUA

Peter Howell and Stuart Rosen
University College London

The voiceless affricate/fricate distinction has figured prominently in current theorizing about speech perception. Two classes of perceptual theory have drawn support from it. First are the natural auditory sensitivities theories in which phonemic categories are thought to be based on simple acoustic properties - the rise time of the frication noise in the case of voiceless affricate/fricative (Cutting & Rosner, 1974; Stevens, 1981). Second are the articulatory referential theories which propose that information about how sounds are produced is used during perception (Dorman, Raphael & Liberman 1976).

Evidence for the view that special auditory sensitivities exist for the perception of rise time comes from studies of categorical perception. Cutting and Rosner (1974) reported that a voiceless affricate/fricative contrast varying in frication rise time and duration was categorically perceived with a category boundary at 40 ms. More importantly, a non-speech continuum consisting of sawtooth stimuli varying essentially in rise time alone was also categorically perceived and the boundary occurred at about the same value of rise time as it did with the voiceless affricate/fricative continuum. These results were though to show that rise times of about 40 ms served as a natural, auditorily-determined boundary which was used in speech to achieve a separation of affricates from fricatives (Stevens, 1981). In a series of studies, we have shown that this interpretation is not tenable. For one thing, a boundary of 40 ms does not distinguish affricates from fricatives in real speech (Howell & Rosen, 1983a). Also, neither the non-speech (Rosen & Howell, 1981, 1983) nor speech (Howell & Rosen, 1984) continua are perceived categorically. There is no evidence for a natural boundary in either case.

An explanation of the perception of the affricate/fricative distinction in terms of a natural sensitivity for rise time implies that distinguishing such sounds is a relatively simple auditory process. Others have demonstrated that a combination of several cues is necessary and that rise time alone will not suffice. Dorman, Raphael, and Isenberg (1980), for example, reported that perception of the affricate/fricative contrast depends on the vocalic portion of the utterance, the duration of the closure interval, the presence or absence of a release burst, and the duration of the fricative noise, in addition to the rise time of frication. These results demonstrate that telling affricates from fricatives is more complex than the natural sensitivities accounts allow. Given that there are a number of cues which interact, there are several hypotheses concerning the way this happens. Delgutte (1982) has argued that two of these cues may interact at the auditory level. Others propose that cues are combined by the listener's use of articulatory knowledge.

An important piece of support for articulatory-referential explanations of the perception of the affricate/fricative distinction is the finding of Dorman et al. (1976), replicated by Repp, Liberman, Eccardt, and Pesetsky (1978). Both studies showed that a longer period of silence before a given duration of noise was needed for an affricate to be reported when the test item was preceded by an utterance spoken at a fast rate than when the precursive phrase was spoken slowly. These data are odd in that it might be supposed that as a sentence is spoken faster, both silence and frication would decrease in direct proportion. Test items preceded by speech spoken at a fast rate ought then to require less silence (and, in fact, noise) for an affricate to be reported.

Repp et al. note that Gay (1978) showed that the duration of all intervals is not reduced equally as speech is spoken at faster rates. The duration of the silence in plosives was affected less than the duration of the surrounding vocalic intervals. Repp et al. argue, following a suggestion in Dorman et al. (1976), that this might explain the anomalous result. They assumed that, as speaking rate increased, the duration of the silent gap associated with the affricate would reduce less than the duration of the fricative noise, as did the silent gap relative to the vowel in Gay's (1978) study. For each of the continua that Repp et al. tested, noise duration was constant and so the noise would take up proportionately more of the sentence as speech rate was increased. The effectively longer noise duration would bias the listener to hearing a fricative, since longer noise durations normally signal fricatives (Gerstman, 1957). Thus, subjects would need proportionately more silence (occurring only with affricates) to offset this bias. In other words, an effective increase in noise duration because of increased speaking rate must be offset by an even bigger increase in the duration of silence that precedes the noise.

Repp et al. consider that such an explanation supports articulatory-referential perception, since: "On that assumption, the boundaries would be set not by the number, diversity or temporal distribution of the cues but by a decision that they do (or do not) plausibly specify an articulatory act appropriate for the production of a single phonetic segment. " (p. 622). Though Repp et al. explicitly require articulatory "plausibility" in the stimuli, certain aspects of them are implausible. For instance, it is likely that rise time varies with speech rate, so the constant rise time they used across stimulus sets would make their stimuli implausible as tokens of speech spoken at different rates. Also, frication duration would not stay constant as speech rate varied as in their stimuli, but would probably reduce in duration. Finally, no burst was included in their stimuli, though these almost always occur at the release of an affricate (Howell & Rosen, 1983a; Isenberg, 1978).

There is also a logical problem in the way they have applied their explanation. Though the articulatory-referential account has been presented as an explanation of what a mechanism would make of plausible stimuli, the research on cue combination has been worked the other way round - i.e., what would a mechanism that uses articulatory knowledge make of a stimulus that a vocal tract is unlikely to produce.

The principal problem for Repp et al.'s explanation is that they had no data that indicated how the duration of silence and frication changed in affricates spoken at different rates. Their account requires that silence is reduced less than frication in duration as speech rate increases. In the absence of appropriate data, Repp et al. relied on Gay's data on plosives, as noted above. However, since their argument involves the duration of intervals within affricates (silence and frication), a more sensible comparison would have been between acoustic intervals within the plosives (say, silence and transitions), rather than between the vowel and part of the preceding plosive. Gay does not report measurements on the transitions in detail, but he does indicate that the relationship Repp et al. would require does not hold. He reports that "... transition time was reduced during fast speech, to about the same degree as that for stop consonant closure, some 5-10 ms" (Gay, 1978, p.225).

It is still, of course, possible that silence in affricates is reduced less than other intervals and, consequently, takes up a bigger proportion of the affricate, as Repp et al., suppose. Presumably it was the wish to get more pertinent data that motivated Isenberg (1978) to measure temporal factors in naturally spoken affricates and fricatives. The corpus consisted of the words "ditch" and "dish" spoken in sentence frames at different speech rates. Isenberg measured the duration of the preceding plosive and vowel, the silent interval (when it occurred), and frication duration. These measurements were then expressed as a proportion of the overall sentence duration, a poor measure since rate can change within a sentence: speakers can speak a sentence fast overall, with local parts spoken slowly and vice versa. Ignoring this point, Isenberg would need to show an increase in the proportion of the sentence taken up by silence relative to frication to support Repp et al.'s argument that the perceptual result is explicable in terms of articulation. An interaction should occur between these intervals and speech rate in an analysis of variance. No such interaction occurred. Though Isenberg claimed support for Repp et al.'s argument based on regression lines fitted to individual subject data, the lack of an interaction in the analysis of variance nullifies this conclusion.

Also, there are other data that flatly contradict the relation required by Repp et al., and sought by Isenberg. Maddieson (1980) measured silence and frication duration of intervocalic voiceless affricates and fricatives in Spanish, English and Italian. He reported that the proportion and duration of silence decreased as speaking rate increased.

To summarize, Isenberg's data supports the articulatory plausibility explanation whilst Maddieson's data indicates the need for an alternative explanation. In order to resolve the discrepancy between these two sets of results, we undertook to measure the duration of silence, rise time and overall duration of affricates and fricatives spoken at different rates (Howell & Rosen, 1983b). Measurements were made on the sentences "I saw a chip/ship in the water" spoken by four speakers, three times each at three different rates (slow, medium, and fast).

The critical data to assess Repp et al.'s explanation derive from the affricates. For these sounds, silence, rise time, and overall duration decreased as rate increased. However, at a fast rate silence takes up a smaller proportion relative to frication, whereas Repp et al.'s explanation requires that silence is a bigger proportion. These results are in line with Maddieson's findings and contrary to those of Isenberg. Isenberg's discrepant results may have been obtained because of the sentence contexts he employed: These were "I meant to say talk ditch/dish fast". In these sentences, the final affricate or fricative is followed by a second fricative. The affricate and fricative or the two fricatives would show considerable coarticulation and could easily lead to errors in measuring the duration of frication, which might be rate-dependent.

These measurements demand a reassessment of Repp et al.'s perceptual experiment: they had argued that longer periods of silence are needed for affricates to be perceived when speech rate increases, because this relationship occurs in the articulation of the sounds, and speech sounds are perceived by reference to articulation. Since the relationship between perception and articulation does not hold, some other explanation must apply.

First, however, an experiment equivalent in all crucial respects to that of Repp et al. was conducted in order to confirm their findings. There are essentially two parts of Repp et al.'s experiment: how the perception of burst of noise as affricate or fricative is affected by the duration of the noise and the duration of a preceding period of silence. Second, how perception of these same stimuli is affected when they occur in sentence frames spoken at different rates.

The sentences "Why don't we say chop/shop again?" spoken by one male speaker were recorded at two different speaking rates. Measurements of the fricative noises were made on the "chop" (affricate) and "shop" (fricative) sentences. The "shop" stimulus spoken at a slow rate was employed as the stimulus to be edited to produce the experimental material. A neutral noise duration (average across affricates and fricatives and across the two speech rates) was calculated. This duration (131.3 ms) was imposed on the slow "shop" by excising a medial portion of the noise. Two other stimulus durations were specified and imposed on the stimulus - 20 ms greater and 20 ms less than this. These three stimuli were inserted into both the sentence frames the "shop"s had occurred in (slow and fast). Eleven different stimuli were produced at each rate and for each duration of frication by inserting a period of silence varying between 0 and 100 ms in 10 ms steps.

The sounds were presented in three blocks in random order to eight listeners. Within each block all sounds had the same frication duration. At each frication duration, the eleven sounds with different amounts of preceding silence at both speech rates were presented ten times each. The listener were asked to indicate whether the test item sounded more like "chop" or "shop".

The results are shown in Figure 1. At the top are the data from the sentences spoken at a slow rate and at the bottom at a fast rate. The ordinates are the percentage of "chop" responses and the abscissae the silent gap duration. The points connected together derive from judgments made about stimuli with the same frication duration (short, medium, long). Each curve from each section of the figure shows that affricate report increases as the duration of the silent gap increases. Moreover, for both sentences rates, affricates are more readily reported to have occurred when the frication duration is short.

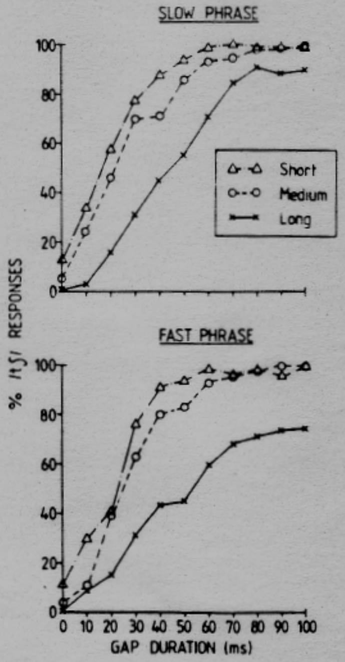

FIGURE 1.

The final feature to note is that affricates are more readily perceived for all frication durations at shorter silent gaps when the sentence is spoken slowly than when spoken fast (the ogives in the top part of the figure are shifted to the left in comparison with those at the bottom). Z scores computed from the maximum likelihood estimates of the phoneme boundaries and the corresponding standard errors, however, show that the difference between slow and fast phrases is only significant when the frication duration is short. Analyses corresponding to these are not reported by Repp et al., though it appears, from the phoneme boundaries reported, that the reverse of this occurred - i.e. there was a bigger effect when frication duration was long. With the exception of this aspect, the results substantiate the findings of Repp et al.

If these relationships are due to processes occurring at an auditory level, then decisions about non-speech analogues might show corresponding differences in the way they are perceived. To check this prediction, the following experiment was conducted.

A non-speech sound (a sawtooth waveform of 100 Hz fundamental frequency) with the same envelope as the sound with the medium duration used in the preceding experiment was substituted for the speech sound. This non-speech sound was inserted into the sentences spoken at two rates and portions of silence varying between 0 and 100 ms (in 10 ms steps) were introduced. These sounds were presented ten times each in random order to eight listeners. The listeners were asked to indicate whether the items sounded like a plucked, or bowed string.

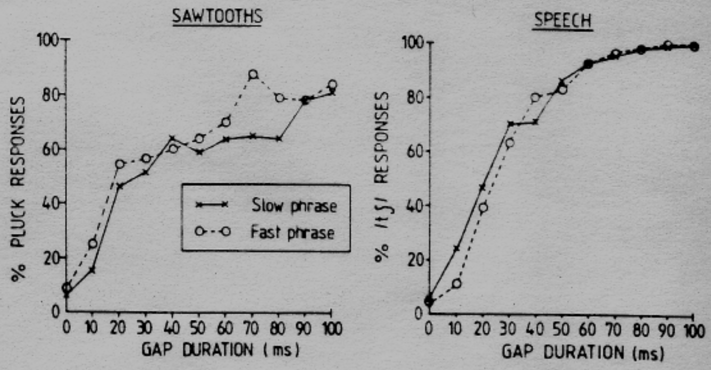

FIGURE 2.

The percentage of pluck responses as a function of silent gap duration are shown in Figure 2 (alongside are shown the corresponding speech data redrawn from Figure 1). The points which are connected together come from conditions which had the sentence at the same rate (the rate can be identified from the symbols and the caption in the inset). It can be seen that at both rates the percentage of pluck responses increases as gap duration increases. The only difference between the speech and non-speech data is that the nonspeech curves are in different order to those with speech, with fewer plucks reported in the slow phrase than the fast. In this case, and unlike the situation with speech, the difference was significant (i.e., Z scores like those computed earlier showed that the category boundaries occurred at a significantly smaller gap duration in the fast phrase [two-tailed]). It is unfortunate that the nonspeech data is only available for one frication duration (this was chosen as a compromise between our own results showing bigger differences at slower rates and those of Repp et al. showing the reverse). Until further data becomes available, it may cautiously be concluded that there is a difference between speech and nonspeech in the rate effect. Clearly, though, perception of pluck/bow and affricate/fricative are both affected similarly by a preceding portion of silence.

The latter data raise as many questions as they answer: first, against the articulatory-referential theory, they show that the effect of gap duration applies to non-speech as well as speech (a similar conclusion follows from the claims of Delgutte, 1982). An auditory explanation of the gap effect might seem appropriate - it is possible that more plucks/affricates are reported for a given rise time when preceded by a silent gap because of the fast-adaptation properties of the auditory nerve (Delgutte, 1982). It would, however, not be possible to account for all of the findings with speech which have been reported by Repp et al., and replicated here on the basis of auditory processes unless additional assumptions are made. In order to account for the rate effects on speech, some forward masking from the diphthongal vowel on the preceding word ("say") would have to be hypothesized. Moreover, the properties of the vowel would have to change with speech rate, so that more masking occurs when the speech is spoken at faster rates, and so that a longer silent period is needed to offset it.

The two obvious candidates for this are vowel fall time and spectral shape - more abrupt falls would produce more masking than gradual ones and energy closer to the frequency region of the following frication should be more influential in masking. Indeed, rough measurements on our sentence frames supports the view that the fall time of the vowel in "say" is more rapid when the vowel is spoken fast than slow. For the sentence frames employed in the test, the fall times of the vowels were 42 and 85 ms measured from oscillograms (employing a similar procedure to Howell and Rosen, 1983a). The same phenomenon can be seen in van Heuven's (1983) intensity displays (his Figure 6). No noticeable difference in spectral shape of the two vowels preceding the affricate/fricative occurred. Thus, it may be that the fall time of the vowel influences judgments about the following sound.

Though this explanation is appealing, it must be qualified, as the vowel and frication noises fall in such different frequency regions and, therefore, little masking would be expected. This qualification does not apply to the sawtooths where, because of the greater spectral similarity between them and the preceding vowel, masking could potentially occur. Yet, in the nonspeech case at the medium frication duration, the results go in the opposite direction to that predicted by the masking explanation (in other words, the results do not confirm the prediction in the condition where the explanation should apply best). Even so, this seems the most informative direction to progress: we plan to measure fall times of a bigger sample of vowels preceding affricates and fricatives at different rates and set up psychoacoustic tests to see whether the envelope of a preceding sound affects rise time perception.

In summary, the data reported here show that the trading of silence and frication at different rates in affricates and fricatives is inconsistent with Repp et al.'s articulatory account. The influence of a preceding gap on perception of a following sound is similar whether the following sound is a burst of frication noise or a sawtooth with the same envelope. This can be accounted for by an auditory explanation: the puzzle that remains is whether and why the functions relating gap duration and pluck/affricate report differ with speech rate.

REFERENCES

1. Cutting, J.E. and Rosner, B.S. (1974). Categories and boundaries in speech and music. Perception & Psychophysics, 16, 564-570.
2. Delgutte, B. (1982). Some correlates of phonetic distinctions at the level of the auditory nerve. In: R. Carlson and B. Granstrøm (Eds.) The representation of Speech in the Peripheral Auditory System. Amsterdam North Holland.
3. Dorman, M.F., Raphael, L.J., and Liberman, A.M. (1976). Further observations on the role of silence in the perception of stop consonants. Haskins Laboratories Status Report, SR-48, 197-207.

4. Dorman, M.F., Raphael, L.J., and Isenberg, D. (1980). Acoustic cues for a fricative-affricate contrast in word-final position. Journal of Phonetics, 8, 397-405.

5. Fitch, H.L., Halwes, T., Erickson, D.M., and Liberman, A.M. (1980). Perceptual equivalence for two acoustic cues for stop consonant manner. Perception & Psychophysics, 27, 343-350.

6. Gay, T. (1978). Effect of speaking rate on vowel formant movement. Journal of the Acoustical Society of America, 63, 223-230.

7. Gerstman, L.J. (1957). Perceptual dimensions for the friction portions of certain speech sounds. Unpublished doctoral dissertation. New York University.

8. Heuven van, V.J. (1983). Rise time and duration of frication noise as perceptual cues in the affricate-fricative contrast in English. In: M. van den Broecke, V. van Heuven, and W. Zonneveld (Eds.), Sound Structures, 141-157. Foris publications, Dordrecht.

9. Howell, P. and Rosen, S. (1983a). Production and perception of rise time in the voiceless affricate/fricate distinction. Journal of the Acoustical Society of America, 73, 976-984.

10. Howell, P. and Rosen, S. (1983b). Closure and frication measurements and perceptual integration of temporal cues for the voiceless affricate/fricative contrast. In: Speech, hearing and language: Work in Progress VCL, 1, 109-117.

11. Howell, P. and Rosen, S. (1984). Natural auditory sensitivities as universal determiners of phonemic contrasts. In: B. Butterworth, B. Comrie, and O. Dahl (Eds.), Explanations of Linguistic Universals, 205-235. The Hague: Mouton.

12. Isenberg, D. (1978). Effect of speaking rate on the relative duration of stop closure and fricative noise. Haskins Laboratories Status Report, SR-55/56, 63-79.

13. Maddieson, I. (1980). Palato-alveolar affricates in several languages. UCLA Working Papers in Phonetics, 51, 120-126.

14. Repp, B.H., Liberman, A.M., Eccardt, T., and Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative and affricate manner. Journal of Experimental Psychology: Human Perception and Performance, 4, 621-637.

15. Rosen, S. and Howell, P. (1981). Plucks and bows are not categorically perceived. Perception and Psychophysics, 30, 156-168.

16. Rosen, S. and Howell, P. (1983). Sinusoidal plucks and bows are not categorically perceived, either. Perception and Psychophysics, 36, 233-236.

17. Stevens, K.N. (1981). Constraints imposed by the auditory system on the properties used to classify speech sounds: Data from phonology, acoustics and psychoacoustics. In: T.F. Myers, J. Laver, and J. Anderson (Eds.), The Cognitive Representation of Speech. Amsterdam: North Holland.

The Psychophysics of Speech Perception

edited by:

M.E.H. Schouten

Institute of Phonetics
University of Utrecht
Utrecht
The Netherlands

1987 **Martinus Nijhoff Publishers**
Dordrecht / Boston / Lancaster
Published in cooperation with NATO Scientific Affairs Division