THE PERCEPTION OF ASPIRATION: A CROSS-LINGUISTIC STUDY

Malou Van Wijk
Universitat Autònoma de Barcelona, Spain

ABSTRACT

The experiment reported here is intended to investigate the strategies applied by speakers in processing speech sounds. It does so by measuring responses of native and non-native listeners to computer-edited stimuli involving natural tokens of English CV sequences with varying degrees of VOT. The specific purpose is to test whether the aspiration period is processed as pertaining to the plosive that induces it or to the vowel that follows it. The results obtained indicate that native speakers tend to integrate the aspiration period in the consonant, but that the presence of aspiration increases perceived vowel length: speech seems to be processed in units that integrate both the consonantal and the vocalic segment. A cross-linguistic analysis seems to confirm the existence of language specific perceptual mechanisms. The findings argue against a segmental analysis and are more in line with a diphone or syllabic model of speech perception.

1. INTRODUCTION

Aspiration is an induced effect of voiceless stops in some languages, such as English, where voiceless plosives are aspirated in syllable initial position in stressed syllables. This paper addresses the question whether the aspiration period is perceptually integrated and segmented by listeners as part of the vowel or as part of the consonant. Aspiration has traditionally been considered part of the consonant. However, it can also be considered the voiceless portion of the following vowel. In acoustic-articulatory terms there is no reason for affiliating aspiration to the consonant rather than to the vowel, since it shows features of both. The fact that the nature of the plosive is responsible for the durational differences in aspiration [1] seems to argue in favour of including aspiration in the consonant. Processes such as changes in vowel quality driven by contextual nasalisation [2] or tonogenesis [3], in turn, indicate that aspiration can be analyzed as pertaining to the vowel. A study on the perceptual integration of the aspiration period in either segment may shed light on the issue. The topic is relevant to segmentation in natural speech processing, speech technology and phonological theory.

The approach of this study rests on the assumptions that, when speakers process speech, 1) they carry out a segmentation and identification of the acoustic events present in the incoming signal and 2) that the units of segmentation are something like the phoneme. In speech perception, lexical access may proceed top-down --with syntactic and semantic knowledge hypothesizing candidates to be compared with aspects of the acoustic signal for verification— or bottom-up --from acoustic signal to lexical item [4]. One of the issues that must be dealt with by any model of bottom-up lexical access is that of segmentation of the signal into phonetic units. The second assumption, that the basic unit of segmentation is something like the phoneme, is rather controversial [2], since features, diphones, syllables and even whole templates like the word have received empirical support as perceptual units. However, evidence that speakers are able to respond to a given sound in experimental situations, and evidence that they are sensitive to mispronounced segments and that they can misperceive --and consequently mispronounce-- a segment in an unfamiliar word in everyday situations indicates that speakers, even if perhaps not primarily, might carry out a segmental type of analysis.

For the understanding of speech processing strategies, it is of great importance to identify the units of speech perception and to study how speakers process an acoustic event such as aspiration: whether they actually use a segmental strategy and, if so, whether aspiration is segmented by the speaker as part of the vowel or as part of the consonant; or whether they process speech in units larger than segments, according to a diphone or syllabic model.

Segmentation is crucial in the field of speech technology. In speech synthesis, the units synthesized and concatenated can be phoneme-like segments, diphones, or larger units. Phoneme-based models of text-to-speech conversion have yielded quite satisfactory results in the past, but considerations of "naturalness" have led to experimentation with larger units. Automatic speech recognition models require accurate specifications on how to segment and process the acoustic events of the incoming signal and on how to classify these phonologically. To expand our knowledge of natural speech perception in general --and the perception of aspiration in particular-- may prove helpful in this respect.

The results of this study may also shed light on phonological issues. Phonological theory has traditionally analysed aspiration as belonging to the consonant that induces it. In many languages, consonantal aspiration, or the phonological feature [+spread glottis], is crucial for the determination of stop categories. If evidence is found for affiliating the aspiration period to the vowel segment, phonological notations might require revision. If no evidence is found for the perceptual integration of aspiration in either the following vowel or the preceding consonant, then segmental phonological models may require revision.

Most theories on syllable structure deny the influence of initial consonants on the following vowel. The Nucleus and Coda are claimed to form a constituent, the Rhyme. Evidence of an Onset-Rhyme structure of English syllables has been found in the fields of stress assignment, speech errors and word games, as well as in distributional constraints [6, 7, 8]. Work on co-occurrence restrictions between the Onset and the following phonemes in a syllable [9] seems to corroborate that the Nucleus and the Coda form a unit in a way the Onset and Nucleus do not. To test the relevance of aspiration to perceived vowel duration may prove revealing in this aspect. If consonantally induced aspiration is found to affect perceived vowel length, it can be considered evidence in support of a dependency relation between onset and nucleus, for which phonological theories will have to make provision.

Extensive research [10,11] has been carried out on the perception of CV sequences and the role of a large range of acoustic cues such as aspiration (or +VOT), yielding extremely interesting results. However, most of the experiments in this area involve synthetic speech (and even non-speech) stimuli. Information from an experiment involving natural (although slightly edited) speech may prove revealing.

2. METHOD

The hypotheses underlying the experiment were that if the aspiration period is considered part of the consonant, then it will not be perceptually integrated in the duration of the vowel (Hypothesis 1); in contrast, if the aspiration period is considered part of the vowel, it will be integrated in the duration of the vowel (Hypothesis 2).

In order to test the above hypotheses, subjects were presented computer-edited natural tokens of [ChVC] and [CVC] sequences which formed nonsense monosyllables. The task required of the speakers was to produce pairwise judgements on vowel duration. The test stimuli corresponded to either Type A or Type B and each stimulus consisted of two contrasting members, as shown in Figure 1.

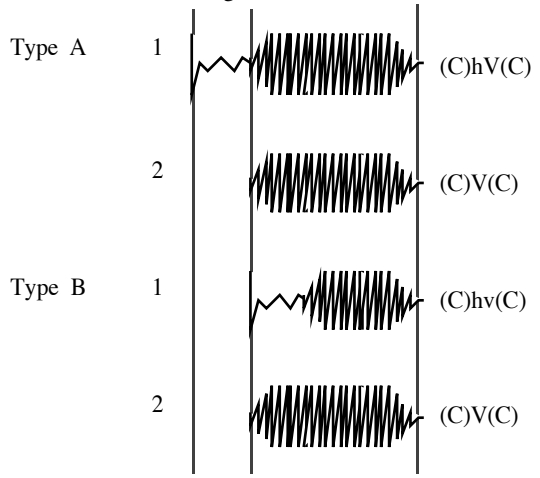

Figure 1. Basic patterns of test stimuli: C = consonant; h = aspiration period; V = voiced vowel, full duration; v = voiced vowel, reduced duration.

A native English speaker provided the raw material for the construction of the stimuli. Each stimulus consisted of two members, which were constructed according to the above patterns from a single [ChVC] token: for 2A and 2B, the aspiration period was segmented and deleted; for 1B, a central portion of the vowel (equalling the duration of the aspiration period) was deleted. Thus, each stimulus contrasted two sequences which were identical but for presence vs absence of aspiration and/or vowel duration. All other variables -- including the order of presentation of the sequences-- were controlled.

Most of the stimuli carried the initial consonant [t], and the vowel [a:], but a number of stimuli were included to study the effect of degree of aspiration, place of articulation of the initial consonant, lack of consonantal context, vowel quality and vowel quantity. The final consonant in all sequences is [f].

The total amount of stimuli was 40: 32 test-stimuli and 8 control-stimuli. They were presented on tape to 35 native speakers of English and 16 native speakers of Spanish or Catalan with no —or very little— knowledge of English. The subjects were instructed to mark 1, 0 or 2 on an answer sheet according to whether they perceived the first, none, or the second vowel in each stimulus as longer.

According to Hypothesis 1 we would expect a majority of 0-responses for stimuli of type A and according to Hypothesis 2 a majority of 0-responses for stimuli of type B. However, the number of 0-responses is misleading, since it includes both cases of definite perception of two vowel durations as equal and cases of indecision. If any decision is taken by a subject, this will be reflected in the 1- and 2-responses. Thus, with a hypothesis "X = Y", our expectation will not be that of 100% 0-responses, but rather that of an equal distribution beween 1- and 2-responses, whatever the number of 0-responses may be. Therefore, if ratings 1 and 2 are not equal, if they are statistically significantly different, the hypothesis will be falsified: $X \neq Y$. In view of these considerations, the above hypotheses yield the following predictions (cf. Fig. 2):

Hypothesis 1: (Asp.= C)	Hypothesis 2 (Asp.= V)
P1: In A, majority of 0-responses	P4: In B, majority of 0-responses
P2: In A, 1-responses = 2-responses	P5: In B, 1-responses = 2-responses
P3: In B, 1-responses < 2-responses	P6: In A, 1-responses > 2-responses

3. NATIVE SUBJECTS

3.1. Results

The responses to the basic stimuli contrasting presence vs absence of aspiration are shown in Figure 2. Figure 2 shows a majority of 0-responses to stimuli of type A: Prediction 1 is borne out in support of Hypothesis 1. However, Prediction 1 is related to Prediction 2 in the sense that, if 1A = 2A, then 1-responses must equal 2-responses. The difference between 1- and 2-responses in A, however, is statistically significant: the vowel in the first member of the stimulus is perceived as longer than the vowel in the second member. This implies that presence of aspiration lengthens perceived vowel duration. Thus Prediction 2 is not borne out, refuting Hypothesis 1, whereas Prediction 6 is borne out, supporting Hypothesis 2.

In stimuli Type B, most speakers judged the second vowel to be longer than the first . The data corroborate Prediction 3 and support Hypothesis 1. Hypothesis 2, in turn, is refuted, since Predictions 4 and 5 fail to be borne out by the data. However, the rather high number of 0-responses indicates that presence of aspiration lengthens perceived vowel duration, although not to the same extent as if the acoustical duration of aspiration were that of a voiced vowel, since in that case we would expect a majority of 0-responses, and an equal number of 1- and 2-responses.

The data in Figure 2 yield contradictory results. There is more support for Hypothesis 1, since Predictions 1 and 3 are borne out while only Prediction 6 lends support to Hypothesis 2. Hence, both hypotheses are partially refuted and may have to be restated.

3.2. Discussion

The data provide no conclusive evidence in favour of either a [Ch-V] or a [C-hV] segmentation. However, the presence of aspiration is found to influence the perceived duration of the following vowel: a vowel is perceived as longer when it is preceded by a period of aspiration than when it is not. The perceived lengthening effect of the aspiration period on the

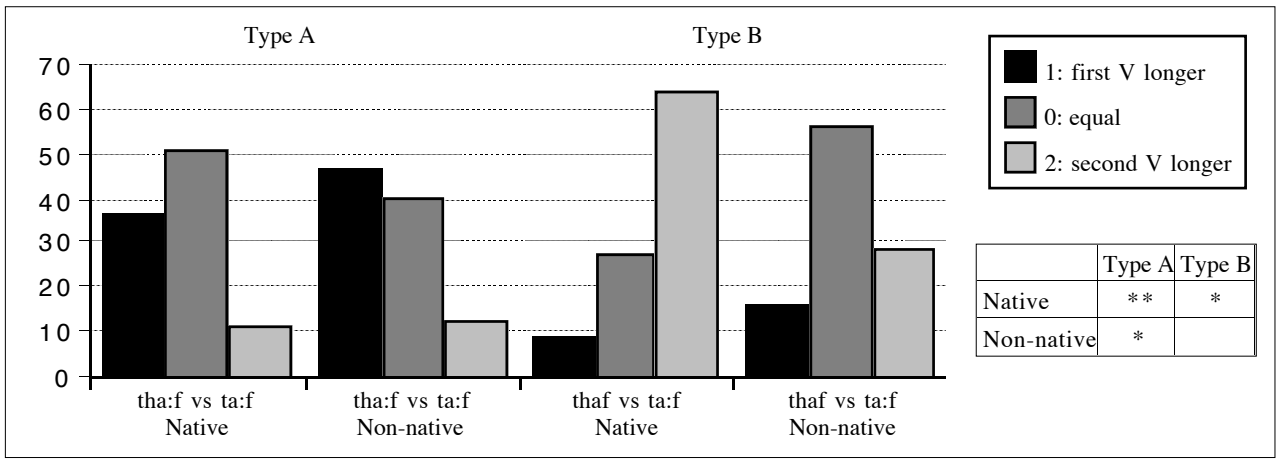

Figure 2. Percentage of identification as 1, 0 or 2 of stimuli presenting sequences with and without aspiration. Stimuli type A and B. Total number of responses for each stimulus: 70 from native subjects: 32 from non-native subjects. Significant differences between 1- and 2- responses for each stimulus type are given in the table at the right of the graphs (Scheffé, *p <0.05; **p <0.001)

following vowel, however, does not equal that of a voiced vowel period of the same duration. Aspiration is found to lengthen perceived vowel duration by a factor inferior to the acoustic duration of the aspiration period.

It does not make any difference whether we postulate a segmentation before or after the aspiration period, since weaker versions of Hypothesis 1 —that aspiration is processed as part of the consonant, but lengthens the perceived duration of the following vowel— and of Hypothesis 2 —that aspiration is processed as part of the vowel, but does not contribute to increased vowel length as if it were a voiced vowel period— make the same predictions

A comparison of the results for the basic stimuli to stimuli involving a 50% reduction of the aspiration period show that the effect of aspiration on the duration of the vowel is not additional: presence of aspiration does not increase perceived vowel duration by a constant factor, since a decrease in degree of aspiration causes a decrease in perceived vowel duration. Rather, the effect of aspiration on perceived vowel length is proportional to the degree of aspiration.

The results obtained for stimuli involving initial stops which differ in place of articulation (/p/ vs /t/ vs /k/) paralleled the results for different degrees of aspiration: the further back the place of articulation of the stop —and consequently the longer the aspiration period induced [1]— the greater the increase in perceived vowel duration.

Differences of vowel height and vowel length cause no systematic variation in the effect of aspiration on perceived vowel duration, since similar results were obtained for stimuli involving high vs low as well as long vs short vowels.

Processing of consonantal aspiration was found to parallel that of /h/-aspiration, since almost identical results were obtained for stimuli with ([tha:f] vs [ta:f]) and without ([ha:] vs [a:]) a consonantal context. Thus the aspiration period cueing the phoneme /h/ influences perceived vowel duration in the same way as an aspiration period induced by a preceding plosive.

The findings show that the aspiration period proportionally lengthens perceived vowel duration. Since the degree of aspiration depends on the place of articulation of the plosive which induces it, the findings suggest that speakers process speech in units that integrate both the consonantal and the vocalic segment. The results obtained in this study, therefore, argue against a segmental analysis and are more in line with a diphone or syllabic model of speech perception.

4 NON-NATIVE SUBJECTS

4.1. Results

Figure 3 shows a clear difference in the distribution of responses from native as compared to non-native subjects. Whereas the results were contradictory for the native group, they are quite regular and conclusive for the non-native group. As in the native group, in stimuli type A the number of 1- and 2-responses were significantly different for the Spanish speakers. In stimuli type B, however, unlike the results for the English speakers, the number of 1- and 2-responses were not significantly different in the case of the Spanish speakers.

Hypothesis 1 can be rejected for the non-native group, as none of its predictions are borne out by the data, whereas all predictions of Hypothesis 2 are sustained, thus yielding conclusive evidence for a C—hV segmentation. This finding is corroborated by responses obtained for stimuli which feature only half the aspiration period, for which the number of 1-responses in both stimulus types decreased considerably.

4.2. Discussion

We have found empirical evidence for different processing strategies applied by native and non-native speakers to Stop + Vowel sequences: English speakers show a tendency to perceptually integrate the aspiration period in the consonant, whereas Spanish speakers quite consistently integrate the aspiration period in the duration of the vowel. These different processing strategies can be accounted for by the fact that neither aspiration nor vowel duration are distinctive features in Spanish or Catalan. The results hence confirm that speakers interpret the acoustic signal in terms of their L1 phonological system. Finally, unlike the data for the native group, the data for the non-native group do not provide evidence —either positive or negative— for a unit of speech perception larger than a segment.

5. GENERAL DISCUSSION

The results described in 3.2. parallel those obtained by Kozhevnikova and Chistovich [12]. Their measurements of the reaction time for the recognition of consonants and vowels in CV-syllables revealed that the basic units of speech perception are the phonemes (cf. our results that native speakers perceptually integrate the aspiration period in the consonant), but that reaction times for the recognition of the vowel depend on the characteristics of the preceding consonant (cf. our results that perceived vowel duration is increased in proportion to the aspiration period) Our results hence support their claim that, although the basic unit of perceptual decision may be the phoneme, the segment of speech to be considered in order to define the phoneme is at least a CV segment. Still, the question remains why aspiration lengthens perceived vowel duration. The finding seems to indicate that speakers expect a shorter vowel after a (longer) aspiration period. Auditory theory [13]might account for this in terms of the temporal organisation of acoustic events within the syllable. Although the results do not contradict the general claims of gestural theory [14], it is however not clear why a motor command for a CV sequence should involve a shorter vowel when following an aspirated stop. In general, however, our results confirm the claim [15, 16] that phonemes may well be real units at some level at which speakers store speech, but that they may not be the units used to encode and decode speech.

The crosslinguistic differrence observed can be accounted for by the claim that the process of native language acquisition [16] causes a general desensitivation to phonetic contrast not exploited in that language. For the non-native subjects in this study, aspiration does not cue any characteristic of the consonant and is hence segmented differently than by native subjects. Similarly, in stimuli type B, which feature a long vs a short vowel (recognised as such by the native speakers), non-native speakers, for whom vowel quantity is not distinctive, fail to perceive this contrast.

In conclusion , speakers can and do attend to segments when they process speech, but even when their attention is specifically focussed on a segment, as in the experiment reported here, the influence from flanking segments is such that it alters their perception. We have hence obtained strong evidence in favour of a model of speech perception which involves a basic unit larger than the segment.

ACKNOWLEDGEMENTS

Part of this research has been funded by grant DGICYT: PB96-1158 awarde by the Ministry of Education, Spain.

REFERENCES

[1] Lisker, L. and A.S. Abramson. 1964. A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements, *Word*, 20, 384-422.

[2] Hombert, J.M., J. Ohala and W. Ewan, 1979, Phonetic explanation for the development of tones, *Language*, 55, 37-58.

[3] Cohn, A.C. 1990. Phonetic and Phonological Rules of Nasalization. *UCLA Working Papers in Phonetics*, 76.

[4] Klatt, D.H. 1979. Speech perception: a model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics* , 7, 279-312.

[5] Ladefoged, P. 1998. The international Phonetic Alphabet as a hierarchical feature system. In S. Axmaker, A. Jaisser and H. Singmaster, eds. *Proceedings of the 14th Meeting of the Berkeley Linguistics Society*, Berkeley, Berkeley Linguistics Society, 124-141

[6] Fowler, C.A. 1987. Consonant-vowel cohesiveness in speech production as revealed by initial and final consonant exchanges. *Speech Communication*, 6:231-244

[7] Hocket, C.F. 1973. Where the tongue slip, there slip I. In *Speech errors as linguistic evidence*, V. Fromkin, ed. The Hague, Mouton: 93-119.

[8] Clements, G.N. and S.J. Keyser, 1983. *A generative theory of the syllable*. Cambridge, MA, MIT Press.

[9] Treiman, R. 1988. Distributional constraints and syllabic structure in English. *Journal of Phonetics*, 16:221-229.

[10] Stevens, K.N. and S.E. Blumstein, 1978. Invariant cues for place of articuoation in stop consonants. *Journal of the Acoustic Society of America*, 64: 1358-1368.

[11] Liberman, A.M. and M. Studdert-Kennedy, 1978. Phonetic perception. In R. Held, H.W. Leibowitz and H.L. Teuber, eds. *Handbook of Sensory Physiology, Vol. VIII: Perception*. New York, Springer Verlag.

[12] Kozhevnikova, V.A. and Chistovich, L.A. 1965. Speech articulation and perception, *Joint Publications Research Service:* 30.543, Springfield, VA, US Department of Commerce, National Technical Information Service.

[13] Pisoni, D.B. 1977. Identification and discrimination of the relative onset of two component tones: Implications for the perception of voicing in stops. *Journal of the Acoustical Society of America*, 61: 1352-1361.

[14] Liberman A.M. and I.G. Mattingly, 1985. The motor theory of speech perception revised. *Cognition* 21: 1-36.

[15] Ohala, J.J. 1986. Phonological evidence for top-down processing. In J.S. Perkell and D.H. Klatt, eds. *Invariance and variability in speech processes*, Hillsdale, NJ, Lawrence Erlbaum Associates: 386-401.

[16] Jusczyk, P.W. 1986. Toward a model of the development of speech perception. In J.S. Perkell and D.H. Klatt, eds. *Invariance and variability in speechprocesses*, Hillsdale, NJ, Lawrence Erlbaum Associates: 1-35.