COMPLEX CONTROL OF SIMPLE DECISIONS IN THE PERCEPTION OF VOWEL LENGTH

Sieb G. Nooteboom, Institute for Perception Research, Eindhoven Netherlands

Introduction

Measurable vowel durations in connected speech are influenced by many different factors. Well-known examples are the effects on the durations of stressed vowels of (1) the postvocalic consonants, (2) the position of the syllable in the word, (3) the syntactic position of the word in the sentence, (4) the presence or absence of a speech pause following the syllable to which the vowel belongs, (5) overall speech rate. Despite the wide variability in acoustic durations due to the combined effects of these and other factors, in many languages vowel durations are contrastive cues to vowel phoneme perception. The fact that the shortest durations of phonemically long vowels can be considerably shorter than the longest durations of phonemically short vowels does apparently not prevent listeners from making correct decisions as to perceived phonemic vowel length. Let us examine how a decision strategy might be organized in order to accomplish this.

A hypothesized decision strategy

We assume, in terms of signal detection theory, that each acoustic vowel duration is represented on an internal stimulus strength axis with an uncertainty that is equal for all vowel durations, at least within the limited range of durations in which we are interested. This uncertainty, due the effects of sensation noise, gives rise to a Gaussian distribution of stimulus strength values for repetitions of each particular vowel duration. We further assume that identification of vowel length is so organized that each individual stimulus strength value X, derived from a single vowel duration, is compared to an internal criterion C on the stimulus strength axis. All X lower than C are identified as short vowel phonemes, all X higher than C are identified as long vowel phonemes. The criterion C is assumed here to be noiseless so that all uncertainty in phonemic decisions has to be caused by the effects of sensation noise. Due to the Gaussian form of the stimulus strength distributions, the probability distribution of short vowel decisions over a set of acoustic vowel durations, and of course the complementary probability distribution of long vowel decisions,

will have the form of a cumulative normal distribution with a mean μ (the phoneme boundary) reflecting the position of the internal criterion C on the stimulus strength axis, and a standard deviation σ reflecting the effect of sensation noise.

Although the internal criterion C is supposed to be noiseless, this does not imply that it has always the same position on the stimulus strength axis. We assume that the listener can move the internal criterion up and down the stimulus strength axis, adjusting its position in order to optimize the chance of correct recognition. Imagine that a listener perceives a vowel segment and is not sure whether the phoneme intended by the speaker was a long or a short vowel. He may then take into account that the vowel concerned is clearly stressed, is followed by a voiceless plosive, is in the final syllable of a word which is immediately followed by a major syntactic boundary, not accompanied by a speech pause, and that the speech rate is slightly faster than normal. Our listener knows from experience that in these conditions a short vowel would have had a stimulus strength value A and a long vowel a stimulus strength value B. He therefore places his internal criterion C in the middle between A and B, in this way optimizing his chance of a correct decision on perceived vowel length.

Of course, such a decision strategy implies that listeners have an extensive and detailed knowledge of the temporal regularities of speech and are able to apply this knowledge very rapidly, so rapidly in fact that they are not aware of doing so. They are even unaware of having this knowledge. We cannot, therefore, test the proposed theory by asking listeners what they do. We need another kind of test. Let us examine a few specific hypotheses that may be derived from the theory and see whether they are corroborated by experimental data.

The effect of sensation noise

If our theory is correct we could measure the accuracy of auditory representation in a vowel length identification task, by determining the σ of the cumulative normal distribution. Of course, in psychoacoustics the accuracy of auditory representation is generally expressed in terms of a differential threshold measured in a binary forced choice comparison task, involving two stimuli per decision. Our theory predicts that the accuracy is equal in both tasks, because there is no reason to suppose that the effect of

sensation noise would be different. In testing this hypothesis, however, we must take into account that in a binary forced choice task involving two stimuli the stimulus separation needed to obtain a given level of performance, for example the 75 % level, is $\sqrt{2}$ greater than the stimulus separation needed to obtain the same level of performance in a similar task involving only one stimulus per decision (Green and Swets, 1966, 68). Because the σ of a cumulative normal distribution is almost $\sqrt{2}$ times the 75 % differential threshold, we may compare the σ measured in a binary forced choice identification task directly to 75 % differential thresholds measured in a comparison task.

In a number of identification tests on the effect of acoustic vowel duration on the distinction between Dutch short /ɑ/ and long /a:/ in different speech contexts, ranging from isolated vowels to vowels embedded in full sentences, we have found σ's averaged over groups of at least 10 listeners for each context condition, between 10 ms (for vowels in isolation) and 3 ms (for one particular full-sentence condition). In most conditions σ's were in the order of 6 ms. Phoneme boundaries ranged from 75 to 100 ms. These σ's are within the range of differential thresholds of sound duration reported in the literature for sounds with approximately the same durations (Lehiste, 1970; Nooteboom and Doodeman, 1978). There is an unpredicted and clear tendency for the σ's to decrease when more speech context becomes available to the listeners. If we stick to our assumption that, within each particular context, the internal criterion C is noiseless, this would mean that the effect of sensation noise is context dependent. If one would prefer to assume that the effect of sensation noise is independent of speech context, one would have to abandon the idea of a noiseless criterion, and assume that the internal criterion shows less uncertainty with increasing embeddedness of the vowel segment.

Moving the internal criterion up and down

Let us now see what happens to the phoneme boundary, being the measurable reflection of the assumed internal criterion C in the decision strategy, when we change the speech context. The proposed strategy implies that, when the speech context changes in such a way that the expected durations of short and long vowel phonemes change, the internal criterion will move up and down accordingly. This prediction has been tested for changes in speech context related to (1) the postvocalic consonant, (2) the position of the syllable in the word, (3) the syntactic position of the word in the sentence, (4) the presence or absence of a speech pause after the syllable, (5) the overall speech rate. The experimental design was very similar in all cases and has been described in detail elsewhere (Nooteboom and Doodeman, 1978). All experiments were limited to the Dutch /ɑ/ - /a:/ opposition. It should be noted that in natural speech these two vowels are distinguished not only by their relative durations but also by their spectral properties, which were kept constant in the experiments. All experiments involved at least 10 subjects. Let us briefly review the results.

- The postvocalic consonant

  Vowel segments followed by a speech pause have generally a greater duration than those followed by a consonant. The amount of shortening caused by the postvocalic consonant depends on the nature of the consonant. For example, plosives tend to shorten the preceding vowel more than do fricatives. This is valid for both short and long vowel phonemes. Consequently the optimal position of the internal criterion C will be at a lower stimulus strength value for vowels followed by a fricative than for vowels in isolation, and at a still lower value for vowels followed by a voiceless plosive. We therefore can predict that the phoneme boundary between /ɑ/ and /a:/ in isolation lies at a greater duration than the same phoneme boundary measured before /s/, which again lies at a greater duration than the one measured before /t/. This prediction is corroborated by the data. We find the following phoneme boundaries, estimated from probability distributions in a vowel length identification test by fitting cumulative normal distributions (sd stands for the standard deviation over the subjects, and is not to be confused with the earlier discussed σ):

In isolation	100 ms	(sd 8.4 ms)
Before /s/	97 ms	(sd 6.7 ms)
Before /t/	91 ms	(sd 6.9 ms)

- The position in the word

  Both short and long Dutch vowels bearing lexical stress tend to become shorter with increasing number of unstressed syllables following in the word (Nooteboom, 1973). Thus we predict that the phoneme boundary will shift towards shorter durations when more

unstressed syllables are added to the word. This has been tested with nonsense words in isolation, in which the first, stressed, syllable contained the test vowel segment, followed by intervocalic /t/, and the number of unstressed syllables was 0, 1, 2 or 3. Phoneme boundaries and standard deviations over the subjects were:

0 unstr. syll. 91 ms (sd 6.9 ms)
1 unstr. syll. 88 ms (sd 5.8 ms)
2 unstr. syll. 85 ms (sd 5.5 ms)
3 unstr. syll. 83 ms (sd 4.3 ms)

Differences between the last three conditions might have been induced by perceived changes in the second syllable.

- The syntactic position of the word

Durations of Dutch short and long vowels in monosyllabic words vary systematically with syntactic position of the word, notably with the type of syntactic boundary immediately following the word. In one experiment we measured phoneme boundaries between /ɑ/ and /a:/ in a monosyllable /tVk/ embedded in 5 different test utterances, each with a different syntactic structure. These test utterances were obtained from sentences spoken with the long vowel /a:/ in the test segment slot, and had normal rhythm and intonation. This original vowel /a:/ had durations ranging from 150 ms in one utterance to 190 ms in another. In each spoken sentence the original vowel segment was excised and replaced by one of a set of test segments differing in acoustic duration. Phoneme boundaries were assessed in each of the 5 test utterances for each of 12 subjects. They ranged from 76 ms, for the test utterance with an original /a:/- duration of 150 ms, to 100 ms, for the test utterance with an original /a:/-duration of 190 ms. Calculating the product moment correlation of the original /a:/-durations in each of the 5 test utterances with all 12 phoneme boundaries in each of these utterances gave $r = 0.83$ ($p < 0.001$). Apparently the /a:/-durations as originally spoken in these test utterances are to a fair extent controlled by the same factors as the phoneme boundaries in the identification test. These factors are either to be looked for in the syntactic structures of the sentences, as intended by the speaker and perceived by the listeners, or, more probably, in the prosodic structures of the sentence realizations, which, of course, are partly determined by the syntactic structures.

- The prepausal position of the syllable

Durations of Dutch short and long vowels in monosyllabic words are considerably longer when the word is in prepausal position than when it is not, other things being equal. Thus we may predict that the phoneme boundary in an embedded monosyllable will shift towards a greater duration when we insert a speech pause immediately after the monosyllable. This has been tested by inserting acoustic silent intervals with durations of 0, 100, 200, and 800 ms immediately after the test syllable /tVk/ embedded in a test utterance, and measuring the phoneme boundary in each of these conditions. In addition, the probability of speech pause perception has been measured independently. It was found that the phoneme boundary increased from 79 ms for a silent interval of 0 ms, to 94 ms for a silent interval of 800 ms, and that the phoneme boundaries in all test conditions were accurately predicted by

$$pb = 79 + 15P_{spp}$$

in which pb is the phoneme boundary in ms, and P_{spp} is the probability of speech pause perception. The probability distribution of speech pause perception over the durations of silent intervals follows an exponential function with a time constant of 200 ms. One possible interpretation of these results is that the listeners employed in this experiment two discrete internal criteria for vowel length identification, one for the prepausal and one for the non-prepausal condition. The gradual increase of the mean phoneme boundary value could be entirely due to the gradual increase in the probability of speech pause perception.

- The effect of speech rate

The effect of speech rate was assessed in a listening experiment employing a computer-controlled channel vocoder in order to vary overall speech rate of a test utterance. In this test utterance the syllable /tVk/ was in final position and the speech rate of all speech material preceding the test vowel segment was 0.67, 1, or 1.5 times normal. Phoneme boundaries were 81, 95 and 102 ms respectively, showing a partial adjustment to speech rate.

Conclusions

The proposed decision strategy for the disambiguation of vowel length is confirmed in all experimental tests. The effect of sensation noise does not conflict with what would be predicted from

differential thresholds for sound duration, although there is an unexpected tendency towards higher accuracy with increasing amount of embeddedness of the test vowel segment. The effect of speech context on the phoneme boundaries is found to be in the predicted direction in all 5 types of contextual differences which were investigated.

The shifts of the phoneme boundary may seem small, ranging from a few ms to about 25 ms. However, they are generally in the same order of magnitude as contextual effects on the durations of short vowels. Analogous to the tendency towards higher within-subject accuracy with increasing amount of embeddedness, we find less inter-subject variation with increasing amount of embeddedness, suggesting that the position of the internal criterion C becomes more and more constrained by inter-personal factors when more speech context becomes available to the listener.

The results support our hypothesis that listeners, whether they know it or not, have an extensive and detailed knowledge of the temporal regularities of speech and actually apply this knowledge rapidly and unknowingly in optimizing their chance of correct decisions on perceived vowel length. This strategy for disambiguation of vowel length may seem an extremely complex and even cumbersome machinery for the communication of a very simple binary contrast. However, given the complexity of the temporal organization of speech, decision strategies in perception have to be complex in order to be efficient.

References

Green, D.M. and J.A. Swets (1966): Signal detection theory and psychophysics, New York: Wiley.

Lehiste, I. (1970): Suprasegmentals, Cambridge, Massachusetts, and London, England: M.I.T. Press.

Nooteboom, S.G. (1972): "The interaction of some intra-syllable and extra-syllable factors on syllable nucleus durations", Institute for Perception Research Annual Progress Report 7, 30-39.

Nooteboom, S.G. (1973): "The perceptual reality of some prosodic durations", JPh 1, 25-45.

Nooteboom, S.G. and G.J.N. Doodeman (1978): "Perception of vowel length in spoken sentences", submitted for publication.