PREDICTING SEGMENT DURATIONS IN TERMS OF A GESTURE THEORY OF SPEECH PRODUCTION

S.E.G. Öhman, S. Zetterlund, L. Nordstrand and O. Engstrand,
Dept. of Linguistics, Uppsala University, Sweden

Theory

We take the basic object of phonetic investigation to be the concrete sound gestalt produced by a speaker in a speech act, i.e. the spoken sentence. And we take the basic problem of our research to be that of explaining the physical structure of the sentence considered, not merely as a complex sound, but as a complex sound used as a vehicle for linguistic communication, or, briefly speaking, we take the problem to be that of explaining the phonetic structure of the sentence.

We explain the sentence phonetically by giving an explicit account of its linguistically functional, physical components, (atomic and compound), and of the methods by which these components are made to form a whole.

An oscillographic or spectrographic record of a spoken sentence does not by itself explain the phonetic structure of the sentence. In such a record, both the (primary) acoustic effects intended by the speaker to form the linguistically functional (atomic or compound) parts of the sentence, and the (secondary) acoustic traces of the speaker's efforts to bring the intended acoustic effects about, will be visible. As a first step in the phonetic explanation of a sentence we therefore try, on the basis of experiment, to distinguish the former acoustic effects from the latter, to define the intended (primary) effects in acoustic terms, and to explain with reference to the physiological properties of the organs of sound production and perception, why the speaker chooses to bring the intended acoustic effects about in just the way he does. In particular, we require detailed explanations of why the (secondary) acoustic traces of these efforts have the acoustic properties that they have.

As an example of this, consider a phonetic part of a sentence that has the form of a voiceless stop consonant such as [t]. It may be assumed that the intended acoustic effect in this case (the [t] per se) is the brief burst of friction noise. The formant transitions that follow this burst (into a following vowel), the

voiceless time interval that precedes it, and the formant transitions (from a preceding vowel) that precede this voiceless interval, are all to be regarded as secondary acoustic traces of the speaker's efforts to bring the burst about. These traces may be explained by showing that a burst of the type in question can only (or at least, most easily) be produced by building up air pressure behind an oral closure at a certain place and by then quickly releasing this pressure in the familiar manner. (A detailed analysis would of course have to make quantitative predictions on the basis of an explicit production and perception model.)

We evaluate an assumption about what does and what does not constitute the intended (primary) acoustic effects, in the total complex sound of a sentence, (1) on the basis of the possibility of explaining convincingly the detailed physical structure of the sentence (including secondary effects), given the physiological constraints on the production and perception mechanisms, and given that the speaker's goal is that of bringing about the assumed primary effects, and (2) on the basis of the depth of insight that the assumed phonetic structure gives us as regards the semantic function of the sentence in the speech act where it was used.

We do not assume that the phonetic structure of a sentence is segmental, nor that it is linear. Experience indicates, on the contrary, that the following picture is better justified:

In any language one operates with a finite inventory of types of atomic acoustic effects. We write E_1, ..., E_n to denote these effect types for a given language (with n such types). Moreover, we use lower case letters (such as e) to denote intended (atomic) acoustic effects, and we write

(1) $\quad e \in E_i \quad$ or $\quad E_i\ (e)$

to say that the atomic acoustic effect e is of type E_i.

Atomic acoustic effects may be combined to form larger acoustic units in either of two ways called coarticulation and sequencing. We write

(2) $\quad e_1 + e_2$

to indicate that the two atomic acoustic effects e_1 and e_2 are coarticulated, which means that they come about (are brought about) simultaneously, or, more accurately, that there is a point in time at which both these effects are heard. And we write

(3) $\quad e_1 \bullet e_2$

to indicate that the two atomic acoustic effects e_1 and e_2 are sequenced, which means that e_1 comes about, whereupon e_2 comes about as soon as can be done. It should be noted that, for all e_i, e_j and e_k the following equalities hold:

(4) $\quad e_i + e_j = e_j + e_i$

(5) $\quad (e_i + e_j) + e_k = e_i + (e_j + e_k)$

(6) $\quad (e_i \bullet e_j) \bullet e_k = e_i \bullet (e_j \bullet e_k)$

(7) $\quad e_i + e_i = e_i \bullet e_i = e_i$

Every language will have special rules according to which certain atomic acoustic effects (specific to that language) can be coarticulated and sequenced. These rules will also allow coarticulation and sequencing of nonatomic (compound) acoustic effects (differently in different languages). If e_1 and/or e_2 are compound effects, the expression $e_1 \bullet e_2$ denotes the sequence in which e_2 develops immediately after the last effect of e_1 has emerged; and $e_1 + e_2$ denotes an effect in which e_1 and e_2 develop simultaneously in such a way that the last effects of e_1 and e_2 coincide in time.

In most languages, we should expect to encounter compound acoustic effects of both of the following forms

(8) $\quad (e_2 + e_5) \bullet e_3$

(9) $\quad e_2 + (e_5 \bullet e_3)$

Here (8) is to be read: the compound effect in which e_2 is coarticulated with e_5, immediately followed (as a whole) by e_3. And (9) is to be read: the compound effect in which e_2 is coarticulated with a compound effect in which the effect e_5 is immediately followed by e_3.

The linearity hypothesis, which we reject, excludes compound effects of the form (9) above.

The acoustic effects are related to articulation as follows. When the speaker says his sentence he knows what acoustic effects he intends to bring about and how they are to be arranged in terms of coarticulation and sequencing. In order to bring these effects about, he makes audible gestures with his organs of speech production, one gesture for each acoustic effect intended, whether atomic

or compound. I.e., the gestures can also be regarded as atomic or compound.

The gestures will be timed and executed in such a manner that (1) the intended acoustic effects come about and (2) no intended acoustic effects are destroyed by the bringing about of other effects.

We hypothesize that in a very considerable number of cases the segmental structure of sentences visible in oscillograms and sound spectrograms and, in particular, the temporal durations of these acoustic segments, can be explained as secondary effects due to the speaker's efforts to bring about the linguistically functional, primary acoustic effects.

The reasoning behind this hypothesis is, among others, this: The linguistically functional, intended acoustic effects are not, in general, required to have any particular duration. They are felt to be complete as soon as they are heard to emerge. A complex acoustic effect in which several atomic effects are coarticulated may, however, require for its execution a compound gesture one part of which is slower than all the others. If several of these gestures are started at about the same time, some of them may be completed earlier than the others in the sense that the effects that they aim at bringing about emerge before the others. To coarticulate all the effects, i.e. make them audible at the same time, the effects that emerge early will have to be maintained for some time while waiting for the remaining effects to materialize. Thus, acoustic segments with quasi-stationary qualities will arise not as a final end of the phonetic action but as a secondary consequence of the effort to reach a certain final end (the simultaneous sounding of the effects in question).

As an example of an alleged phonological contrast that seems eliminable on this paradigm we offer the Swedish contrast [vi:la] [vil:a] (rest, house) which we analyze as

v (stress + i) ● l ● a

v (stress + (i ● l)) ● l ● a

where the stress effect which it takes a relatively long time to produce must be coarticulated with the vowel [i], (thus making the quickly producible [i] long) in the first case, whereas the stress effect is coarticulated with [i ● l] in the second case (thus making the [l] long).

Among the acoustic effect types of most languages there will be certain (relative) pitch levels or compounds (sequences) of such levels. These pitch levels will in general be coarticulated with acoustic effects with the feature [+voice], especially vowels. We therefore expect that vowel duration will be strongly dependent on intonation in most languages.

In what follows some experimental data that have been collected to test this theory will be summarized.

Data

In two experiments (Zetterlund et al. 1978, Engstrand et al. 1978) we used a computer system (ILS) for manipulating prosodic parameters in natural speech to show that listeners consistently tend to overlook systematic durational variations in their identification of certain noun phrases such as compounds and lexicalized phrases. In an identification experiment we presented our informants with various synthesized versions of certain utterances (see Zetterlund et al. 1978, Engstrand et al. 1978) systematically changing fundamental frequency, vowel and consonant durations, and intensity. The responses were consistent to almost one hundred percent: The critical parameter for the listeners' identification of these utterances was F0. Although the acoustic analysis displays great variations in the duration and intensity parameters, our subjects apparently paid no attention to these potential cues in the presence of F0.

On the basis of the theory sketched earlier in this paper, we expect that a large F0 movement between two critical values would tend to space these points further apart in time than a small (or no) F0 change. To test this hypothesis we have in one experiment looked at words involving the Swedish word accent opposition. In bisyllabic accent 2 words in focus position F0 has to be low at the end of the first (stressed) vowel. The second vowel carries the sentence accent which is physically signaled as a high F0 at the beginning of the vowel. Consequently, most of the F0 rise has to take place during the intervening consonant occlusion. The corresponding accent 1 words do not display this upward shift but are characterized by a more or less level F0 contour during the consonant. The interesting thing about this is that the consonant in the accent 2 words seems to be significantly longer than the consonants in the corresponding accent 1 words.

It is known that F0 variations are accompanied by considerable vertical movements of the entire larynx box. Combined electromyographic and larynx movement data that we have collected show that these vertical movements have definite muscular correlates, namely activity in the geniohyoid and sternohyoid muscles for upward and downward movement, respectively. Although the way the vertical movements mechanically affect the tension of the vocal folds is very much open to question, we can state that there is a very high positive correlation between larynx height and F0, and that F0 control involves a delicate coordination between small intrinsic musculature and larger supra- and infrahyoidal muscle masses. And, further, considering the comparatively large mass of the larynx box, it seems rather plausible to assume that its mechanical inertia in combination with the coordinative demands on the muscles should impose restrictions on the velocity with which it can conveniently and accurately be moved.

A further example is this: In an accent 1 word F0 is phonologically required to be low at the beginning of the stressed vowel. If the accent 1 word is preceded in the sentence by a relatively high F0, a downward movement of F0 is observed at the beginning of the accent 1 word sometimes extending into its stressed vowel. The duration of the consonant preceding the stressed vowel is found experimentally strongly to depend on the extent of the pitch drop through the beginning of the accent 1 word.

Finally, in order further to test the theory we have looked at the dispersion of the F0-values at various critical points and found that the standard deviations generally are extremely small, e.g. 2.6 Hz at the last peak of the hu segment in vita huset. We have several more examples of this kind.

Obviously, the most important task must be to establish the critical F0-values at different points in time with greater certainty. The perceptual tolerances that listeners have to deviations in the time-frequency domain should be investigated. We would also like to know what significance the exact shape of the F0 contour between the critical points have to a listener. As a matter of fact, the question whether the entire F0 contour or only some fixed values at certain line-up points relative to supraglottal articulations are the intended and, therefore, phonologically crucial effects produced by the speaker is not yet completely

answered. Pilot experiments encourage us to believe that the latter hypothesis will prove to be true. This would mean, then, that a speaker is given a relatively large amount of freedom to choose ad hoc strategies for passing through the chain of successive critical F0 points. If this is true, a reasonable assumption is that the way transitions are brought about is adapted to fit some anatomical constraints on the larynx. Looking at our data we observe that the slopes of F0 rises and falls are characterized by constancy rather than variation.

References

Engstrand, O., L. Nordstrand and S. Zetterlund: Experiments on the perceptual evaluation of prosodic parameters in compounds and lexicalized phrases. Paper given at The Phonetics Symposium held at the Department of Speech Communication, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, November 9-10, 1978.

Zetterlund, S., L. Nordstrand and O. Engstrand: An experiment on the perceptual evaluation of prosodic parameters for word boundary decision in Swedish. Paper given at The Symposium on the Prosody of the Nordic Languages, Phonetics Laboratory, Department of General Linguistics, Lund University, June 14-16, 1978.

Öhman, S.: Aktuell svensk forskning i fonetik. Tionde sammankomsten för svenskans beskrivning, Uppsala, april 1977. In: S. Eliasson, B. Loman, B. Sigurd, U. Telemann and S. Öhman: Svenskan i modern belysning. Fem översikter från Tionde sammenkomsten för svenskans beskrivning (Ord och stil. Språkvårdssamfundets skrifter 9.) Lund: Studentlitteratur.