The Quantitative Organization of Speech

Christopher Sappok

Institut für Kommunikationswissenschaften, Friedrich-Wilhelms-Universität Bonn, Germany
csa@ifk.uni-bonn.de

Abstract

In the course of listening to an utterance, how can the listener infer how long the utterance is going to be altogether? If this information is available at an early point, it allows to rule out on-line all instances of total utterance structure possible, which do not fit the anticipated window. Hence, decoding effort would be disburdened immensely. The model introduced in this connection gives an integrating view on prosodic surface phenomena such as phrasing/euphony as well as rhythm/isochrony, tracing the temporal organization of different units to a common deeper level representation referred to as the Quantitative Organization of speech (QO).

Index Terms: phrasing, prominence, pauses, temporal structure, implicit prosody, balance principle

1. Introduction

The Call for Proposals of Speech Prosody 2010 states that ”prosodic prominence and phrasing give listeners a window into the mind of the talker”. This is a position paper to model the processing behind this aspect of prosody in detail.

The assumption that the overall duration of an utterance is – somehow – being communicated around its beginning implies that the speaker himself knows it in the first place. And to assume such anticipatory knowledge means to contradict the dictum that speech production is incremental in nature, with look-ahead of one phonological word at most [1].

The present paper summarizes ideas developed exhaustively as part of [2]. The rest of the first section introduces a number of basic considerations. Section 2 lists a set of assumptions more or less controversial with respect to conventional linguistic thought, which for the sake of briefness need to be attributed axiom-like status. In section 3, a model covering selected aspects of speech production is presented. With a representation labeled QO, this model formulates an autonomous link between Linguistic Organization (LO = all aspects of linguistic structure) and Temporal Organization (TO = all aspects of segmentation in the acoustic signal). Section 4 gives an outlook.

As to the order of magnitude of utterance-length anticipated by the speaker and communicated to the hearer, some very general evidence can be called on: [3] observe that metric poetry of all times and cultures is based on a unit named the *3-second-LINE*, deducing that “correspondence between poets in such widely different cultures surely points to an identical neurophysiological mechanism [3].” Given that some universal value for unmarked utterance-length is part of man’s shared knowledge about speech, one bit of information about actual utterance-length communicated early on could be, whether it will be either shorter or longer than that.

Further speculation can be derived from conceptions of the role of symmetry in language-processing, such as the implicit prosody hypothesis (IPH) on silent reading. It states that “[unmarked] prosody is mentally projected onto the written or printed word string. And […] it is then treated as if it were part of the input [4].” Unmarked prosody in this sense results in perceiving quantitatively balanced units, leading to different interpretations of syntactically ambigue expressions such as (1) and (2) (adapted from [5]).

the recently divorced bishop’s daughter (1)
the divorced bishop’s daughter-in-law (2)

Thus, (1) is preferably taken to be about a bishop’s daughter who is divorced, whereas in (2), it’s the bishop who is taken to be divorced. Even though IPH is about silent reading, it allows for the general conclusion that another bit of information concerning utterance-length communicated fairly early in speech could be of the sort: ‘Now, this was half of it – what follows will be of about the same length.’

[5] discusses two principles by which the balance of units may be organized, *multiple branching* vs. *bisection* (Fig. 1), favoring the second. Whereas the considerations of [5] are centered on the phrasal level with respect to cognitive organization, [6] find the same pattern (cf. Fig. 1 b.) in the TO of units on the foot-level and below: “There are [...] approximately 1, 1/2, 1/4, 1/8 relations in the timing of inter-stress intervals, stressed syllables, unstressed syllables and phonemes.”

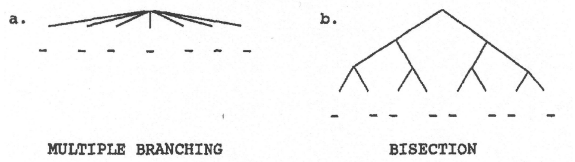

Figure 1: *Possible principles underlying balanced unmarked prosody (reproduced from [5]).*

The general idea of the present approach can now be outlined as follows: If some clearly defined quantitative relation (e.g. bisection) holds all the way from the syllable or even the phoneme level up to the phrasal level, the actual duration of the first few small units would be enough for the listener to project from how long the utterance will be – if also there was information on the scheduled ‘height’ of some sort of ‘prosodic tree’ (as in Fig. 1 b.).

A flaw with strictly symmetry-oriented conceptions is that in practice, prosodic unmarkedness seems to be a question of even quantity more than of even number. For unevenness as an option, some very general evidence can be instanced, e.g. Miller’s *magical number seven* [7] as an elementary determinant of short term memory (STM). Especially interesting I find the fact that in metric poetry of all times and languages, the difference between meters with even and uneven numbers of either ictusses in a verse (cf. Shakespeare’s pentameter) or else verses in a stanza (cf. Dante’s tercet) does not appear to be crucial. I assume that the essential features of metric poetry are definite manifestations of unmarked prosody, or rather unmarked QO – following Lehiste’s premise “that the suprasegmental structure of a language is crystallized, as it were, in the metric structure of its traditional poetry [8]” to universality.

2. Assumptions

The modeling of the role of QO presented in section 3 rests on the following assumptions. (In addition to the sources already cited, key concepts are borrowed from [9], [10], [11].)

2.1. QO is represented on a deeper level.

This statement distinguishes QO from prosody in a wide sense. On the surface level, underlying QO is represented as TO (whereas its surface correlates are not durational only). Speaking of QO means to imply the existence of a cognitive module independent from other aspects of prosody, e.g. *intonational meaning* [1] (whereas those aspects can again affect durational features on the surface).

2.2. QO-processing is not incremental and not serial.

In dealing with "the temporal structure of speech", [9] and [10] unconditionally invoke Levelt's incremental and serial model of speech production [1]. The present approach strictly decouples QO-processing from major aspects of linguistic processing: The main steps take place early on in utterance planning, employing look-ahead of some 3s. The brain areas involved may be located in the right hemisphere, operating in parallel with left hemisphere linguistic areas. The main claim is not about lateralization though, but about differences in underlying principle.

2.3. The principle of LO is hierarchy, the principle of QO is harmony.

Significant here is the idea of non-constituency: in the perspective of QO, intervals of different orders of magnitude are not interrelated with primary respect to the very next (upper or lower) order, but with primary respect to the highest order. Therefore, a metaphor more adequate than the 'tree metaphor' to describe QO is an acoustical one: orders of magnitude are interrelated as fundamental tone and selected overtones are in a harmonious signal. In order to point out the difference with the constituency-based approach of Figure 1, Figure 2 'translates' the notion of harmony into a tree-diagram.

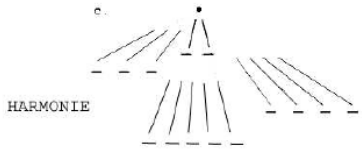

Figure 2: *Tree-fashioned diagram of harmonious relations between events such as the regular intervals of overtones (dashes) with respect to one fundamental interval (dot) (reproduced from [2]).*

2.4. The order of magnitude fundamental to QO is the order of production units.

The term *production unit* is adapted from [11]. The production unit concept is also closely related to the *performance structure* of [9], though avoiding reference to *structure* in its usual hierarchical configuration. Completed production units can be said to correspond to phrases, but the concept is centered on the production process, so the definition is rather: 'what is being planned in one go'. The precept behind this is that speaking is a cyclic procedure with respect to production unit order, which again is determined by STM-span, so in the frequency domain, the range relevant in production unit succession is roughly 1 to 1/4 Hz.

2.5. The operating domain of QO-processing is cohesion.

"By 'cohesion' is meant frequency of co-occurence, semantico-syntactic relationships [...] and syntactic relationship [...]. There is now considerable evidence that speakers seem to organise their speech with reference to such an internal notion of cohesion between the various segments of their utterance [9]." – A similar conception of cohesion as a key cognitive feature is crucial to the modeling of QO. In general, cohesion is conceived of as a continual contour. It can be attributed to any utterance, representing the 'shadow', so to speak, of the whole of LO. QO-processing is modeled as a set of mathematical operations playing upon this contour.

3. Model

The model reconstructs the building-up of specific QO during the production of a single production unit. The steps described lead from the start of planning to the point at which QO is fully developed and transferred to an acoustic representation. The relation between QO and its specific acoustic correlates (i.e. TO) is not treated. The role of QO in speech reception is treated in the closing part of section 3.

Input to QO-processing is cohesion conceived of as a compositional meta-feature. In the perspective of a cognitive QO-module, all of LO is stripped of its qualitative dimension. Analytically, cohesion as a contour $c(x)$ is defined as the result of m contours of layer-specific cohesion – 'shadows' again – additively superimposed (Figure 3).

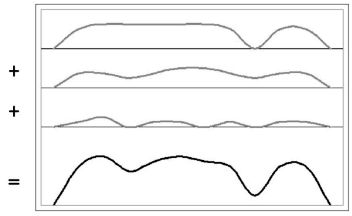

Figure 3: *Compositional cohesion c(x) (black) as the sum of m layer-specific contours $c_1(x)...c_m(x)$ (grey); here: m = 3; no specific utterance implied.*

Cohesion-relevant layers of LO may be connected with all aspects of conceptualization and formulation according to [1]. Keywords in this connection would be: information structure, propositional, syntactic, and phonological structure. Still, the exact number and nature of cohesion-relevant layers is not precise at this point. One problem here is that, with respect to cohesion, mutual dependencies are to be assumed between the layers, whereas additive superimposition implies that there is no redundancy between respective contributions.

How the quantification of structural conditions could be operationalized in detail remains to be developed as well (for an operationalization of solely syntactic cohesion see [12]). One problem here is that, in opposition to the underlying structural conditions, resulting cohesion is not completely determinable – individual or situation-specific biasses in favor of one layer or the other may be assumed. This is taken into account by inserting a weighting factor with each layer-specific structure, which controls its influence on cohesion.

Some basic assumptions about the configuration of layers are already indicated by Figure 3, one about granularity, another about chronology: Layers connected with early stages

of production (e.g. aspects of conceptualization) bear coarse-grained cohesion, later stages bear fine-grained cohesion (e.g. aspects of phonological encoding). This is important for the differentiation between anticipated cohesion and finished cohesion that follows.

The variables involved so far are: $s_1(x)...s_m(x)$ for 'structures 1 to m', with, e.g., s_s standing for a detailed syntactic tree, and $c_1(x)...c_m(x)$ for '(weighted) cohesion 1 to m', adding up to compositional cohesion. As LO has so far been taken to be completely available, this result shall be specified $c_{fin}(x)$ – cohesion finished with respect to the production process. The weights $w_1...w_m$, however, are taken to be constants within a specific situation, with, e.g., all weights except w_s set to 0, which (almost) describes the approach of [12]. The so modeled role of LO is illustrated in the left part of Figure 4.

3.1. Step one: deciding on production unit extension

A fundamental precept of the present model is that one of the very first things a speaker does in 'manufacturing' production unit n, is to hypothesize about its extension with a precision all the way down to about 1/100s. This anticipation value shall be named t'_n – in opposition to a value t_n, which stands for the eventual duration of production unit n in terms of TO (determinable soonest at 'moment' $n+1$). In order to come up with hypothesis t'_n, the speaker takes the information he can get at the beginning of 'moment' n.

On the one hand, there is what shall be called anticipated cohesion $c_{ant}(x)$. $c_{ant}(x)$ is built up like $c_{fin}(x)$, except that it is more vague, given that the planning on the layers appearing high in Figure 3 can be assumed to be somewhat noncommittal and the encoding on the lower layers can be assumed to be only just or not even yet begun. Deciding on production unit sequencing can be seen as horizontal discretization on the basis of $c_{ant}(x)$ and a threshold constant of, say, $y=0.5$. The first $c_{ant}(x)$-minimum lower than this is selected in order to determine t'_n.

On the other hand, there is information about the immediate past relevant to the decision in question. Given that production unit n is preceded by production unit $n-1$, speaker (and listener) know the duration t_{n-1} of the predecessor. Depending on this value, $c_{ant}(x)$ – so far only a matter of anticipated LO – undergoes certain transformation before applying the threshold and deciding on the value to be given t'_n. This transformation is operationalized by equations (3) to (6). They attempt to describe in detail how in deciding on t'_n, a tendency to produce units of similar length is kept.

$$c_{mag}(x)\begin{cases}\left[c_{ant}(x)\right]^{1+\left(q\cdot\frac{x}{p}\right)} & for \quad 0 \le x \le p \\ \left[c_{ant}(x)\right]^{(1+q)-\left(q\cdot\frac{x-p}{p}\right)} & for \quad p \le x \le 2p \\ c_{ant}(x) & else\end{cases} \tag{3}$$

The transformation consists of a continuous 'magnifying effect' on $c_{ant}(x)$ (cf. the magnifying glass in Fig. 4). This effect is governed by two variables: p, defining the point of maximal magnification, and q, defining the maximal extent of magnification. The effect works on an exponential basis, amplifying $c_{ant}(x)>1$ and damping $c_{ant}(x)<1$, all in all resulting in $c_{mag}(x)$. Default values, e.g. relevant for the first production unit in a sequence, are, say, $p=2.5$ and $q=1$.

The influence of this transformation can be illustrated as follows: Given that there are two minima $c_{ant}(2.3)=0.69$ and $c_{ant}(2.9)=0.69$, the default values for p and q would according to (3) lead to $c_{mag}(2.3)=0.49$ and $c_{mag}(2.9)=0.51$. Thus, the first minimum would be selected to determine t'_n, although from the linguistic perspective of c_{ant}, both minima are not small enough to fall under the threshold value. In [2], it is claimed that p and q are variables subject to certain calibration processes. If the preceding production unit would have been long enough to give p a value of, say, 2.8, the second minimum would have been selected. Equations (4) to (6) attempt to describe the underlying calibration mechanism. They are presented here without further discussion.

$$r = t_{n-1} - p_{n-1} \tag{4}$$

$$p_n = \frac{p_{n-1} + (p_{n-1} + r)}{2} \tag{5}$$

$$q_n = \frac{q_{n-1} + \left(1 + \frac{1-|r|}{1+|r|}\right)}{2} \tag{6}$$

(3) to (6) neither claim to work with optimum mathematical operations nor with optimum values. They are not meant to be a precise answer but rather a way of refining the question. Crucial are the similar length tendency via modification of anticipated cohesion and the fact that the variables governing this modification are determined by past events. Hence, if equations (3) to (6) represent shared knowledge, it can be assumed that the speaker and listener of production unit n know the values for p_n and q_n equally well.

3.2. Step two: the 'fanning out' of grids

So far, t'_n has been treated as (anticipated) duration measured in s. Now it will be treated as a frequency-matter, acting as a 'meta-tone' fundamental to the QO of the current production unit. In other words, from specific t'_n a projection is made like a pulse spectrum, with overtones of $2/\ t'_n$, $3/\ t'_n$, $4/\ t'_n$ etc. Hz. At this point, the amplitude of fundamental tone and overtones is not yet specified. In the time domain, this gives the picture indicated by Figure 2: a complete layered system of grids of isochronous intervals, with every interval being associated with an exact (anticipated) duration. This system can be labeled the proto-QO of production unit n.

3.3. Step three: interaction of representations

The interaction of linguistically and quantitatively motivated representations has been addressed in literature as *projection* [5], *transformation* [4], *mediation* [10], or *superimposition* [6]. In the model this interaction takes place in a source-filter configuration. The interacting representations are as source the t_n'-based complete spectrum with 'overtones' and as filter-function cohesion, determining an envelope with 'formants'. (Note that the term cohesion is used vaguely here. In the following, the model is restricted to a reading task context, so that cohesion may be interpreted as $c_{fin}(x)$.)

As $c_{fin}(x)$ is defined as purely linguistic and accordingly 'prosaic' with respect to balance, there would be some divergence between the defining frequencies of 'overtones' and 'formants'. The model presupposes discretization at this stage, in the sense that only one overtone per formant is selected. The result is the QO of production unit n, visualized best in the time domain in the form of corresponding selected interval layers (cf. the grey bars in Fig. 4). Admittedly, this process requires further specification. It is to be pointed out, though, that the number of layer combinations possible in this connection is fairly limited and that there might be an even more limited number of standard combinations.

The next question is, how specific QO may be encoded in the TO of the resulting signal. If the listener reconstructs certain aspects of underlying QO at an early point, he can calculate how long the production unit is meant to become by the speaker. Just as absent fundamental frequency is accessible as a residual tone from the presence of at least two overtones, t'_n gradually becomes accessible to the listener in the course of listening. This would be possible from, say, the first phonetic foot on, given that TO at this point reveals the position of an underlying 'syllable overtone' and a 'foot overtone'.

To make this possible, one last step in production is assumed, the effect of which can be compared to a professional singer adjusting the position of formants to selected overtones in order to achieve maximum resonance. In the present context, to arrive at the TO of the production unit in question, certain aspects of LO such as *"citation" metrical spellout* [1] are altered, durations are adjusted etc. – with respect to cohesion as manifested in the signal. The objective behind this is to give the listener a picture of an envelope by which the underlying harmonics can be reconstructed.

To achieve this, it may only be necessary to effectuate slight shifts of 'formant' position towards overtone position. The resulting form of cohesion shall thus be named $c_{otc}(x)$, i.e. 'overtone centered cohesion'. This representation corresponds to the phonetic plan of [1]. TO itself is manifested in the acoustic correlates of $c_{otc}(x)$: cohesion in the signal $c_{sig}(x)$, whereas perceived cohesion $c_{perc}(x)$ by psychoacoustic effects may be very much the same as $c_{otc}(x)$. The first step in decoding the beginning of a production unit would from this view be to reconstruct $c_{fin}(x)$ and to compare it to actual $c_{otc}(x)$, interpreting the difference with regard to QO and scheduled length respectively.

4. Outlook

Figure 4 summarizes the model. The specific instance of QO indicated can be related to (7):

For the 'benefit of 'mister kite, | 'there will be a 'show tonight | on 'trampoline. || (7)

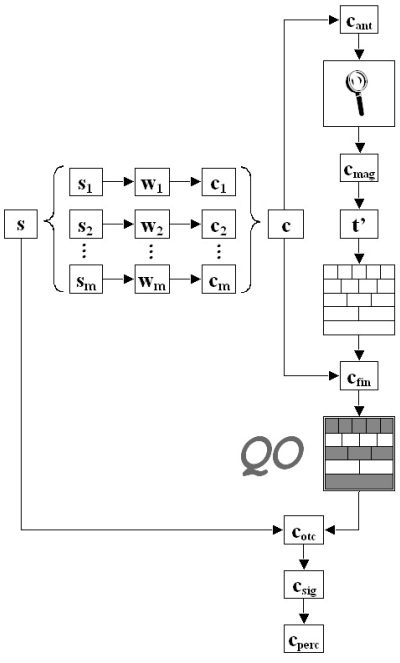

Figure 4: *QO, illustrated here by (7).*

Legend: s = linguistic structure (i.e. LO); s_i = layer-specific structure; w_i = layer-specific weight; c_i = layer-specific cohesion; c = compositional cohesion; c_{ant} = anticipated cohesion; magnifying glass: magnifying effect as described in (3)-(6); c_{mag} = magnified cohesion; t' = anticipated production unit extension; complete system of grids: proto-QO (source); c_{fin} = finished cohesion (filter); selected system of grids: QO. Aspects of TO: c_{otc} = overtone centered cohesion; c_{sig} = cohesion in signal; c_{perc} = perceived cohesion.

The view presented is far from being fully developed. At the present point it allows for an interesting perspective on conceptions like the isochrony hypothesis: The differentiation between stress timed and syllable timed languages would be due to language specific settings concerning the amplitude of the 'formants' relevant in QO-processing. Another conception challenged is that of prosodic hierarchy, as, in terms of QO, prosodic units of different orders of magnitude do not directly constitute each other. Very important is the question of how to verify all the speculation involved. It would be necessary to operationalize the cohesion-conception in full detail. Here the aspect 'frequency of co-occurrence'/statistics needs to be taken into account, which has been neglected so far. Then it should be possible to implement the model in a TTS-System. But the most important question is how to link QO with its acoustic correlates (TO). In a first step reading experiments were conducted, in which each speaker read the same text several times using different production unit sequencing. The data were elicited with selected rhymeless metric poetry as a stimulus. It was once presented in its usual line-configuration and once in prose-like paragraphs. This results in considerable differences in production unit sequencing and syllable prominence that need to be investigated further.

5. References

[1] Levelt, W. J. M., "Speaking – From Intention to Articulation", Cambridge, MA, MIT-Press, 1989.

[2] Sappok, C., "Das deutsche Komma im Spiegel von Sprachdidaktik und Prosodieforschung", Ph.D. dissertation, Leibniz Universität Hannover, Germany, forthcoming.

[3] Turner, F. and Pöppel, E., "The neural lyre: poetic meter, the brain, and time", in Poetry (August), 277-309, 1983.

[4] Fodor, J. D., "Psycholinguistics cannot escape prosody", in Speech Prosody Int. Conf. Proc., Aix-en-Provence, France, 83–88, 2002.

[5] Fodor, J. D., "Learning to Parse?", in Journal of Psycholinguistic Research 27, 285–319, 1998.

[6] Fant, G. and Kruckenberg, A., "On the quantal nature of speech timing" in Speech Acoustics and Phonetics, G. Fant, Dordrecht, Kluwer Academic Publishers, 224-231, 2004.

[7] Miller, G., "The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information", in Psychological Review 63, 81–97, 1956.

[8] Lehiste, I., "Phonetic investigation of metrical structure in orally produced poetry", in Journal of Phonetics 18, 123–133, 1990.

[9] Zellner, B., "Pauses and the temporal structure of speech", in Fundamentals of speech synthesis and speech recognition, E. Keller, Ed., Chichester, Wiley, 41–62, 1994.

[10] Keller, E. and Zellner, B., "Motivations for the prosodic predictive chain", in ESCA/COCOSDA Int. workshop on Speech Synth. Proc., Jenolan Caves, Australia, 137–141, 1998.

[11] Horne, M., Frid, J. and Roll, M., "Timing Restrictions on Prosodic Phrasing", in Nordic Prosody IX Proc., Lund, Sweden, 117–126, 2004.

[12] Abney, S., "Prosodic Structure, Performance Structure and Phrase Structure", in Speech and Natural Language Workshop Proc., Harriman, NY, 425–428, 1992.