Zellner, B. (1994). Pauses and the temporal structure of speech, in E. Keller (Ed.) *Fundamentals of speech synthesis and speech recognition.* (pp. 41-62). Chichester: John Wiley.

Pauses and the Temporal Structure of Speech 3

Brigitte Zellner

Laboratoire d'analyse informatique de la parole (LAIP)
Université de Lausanne, CH-1015 LAUSANNE, Switzerland

Natural-sounding speech synthesis requires close control over the temporal structure of the speech flow. This includes a full predictive scheme for the durational structure and in particuliar the prolongation of final syllables of lexemes as well as for the pausal structure in the utterance. In this chapter, a description of the temporal structure and the summary of the numerous factors that modify it are presented. In the second part, predictive schemes for the temporal structure of speech ("performance structures") are introduced, and their potential for characterising the overall prosodic structure of speech is demonstrated.

Text-to-speech synthesis requires the conversion of phonemic segments into audible speech events. To appear realistic to the human ear, artificial speech must contain natural-sounding vocal inflection, rhythm and stress placement. In other words, speech synthesis requires prosodic features. Temporal phenomena, such as pauses, syllable prolongations and overall timing structure, form an important part of these prosodic aspects of speech. Some such phenomena (like utterance-final syllable prolongations) have been implemented in speech synthesis devices for some time, while others (such as pauses and speech rhythm) are not yet in common use. A fully human-like implementation of these temporal aspects of speech can be expected to lead to important further improvements in speech synthesis by rendering it even more "fluent," more "human-like," and probably also quite a bit more intelligible.

Endowing speech synthesis with prosodic parameters means that intonation, stress, syllabic length and speech rate have to be generated on the basis of textual material. It is therefore important to consider how temporal phenomena occur in human speech, and how they relate to the textual material from which they are generated. At the level of the acoustic;signal, high-level temporal parameters are translated not only into corresponding low-level durational variations, but also into modifications of fundamental frequency and intensity. A second consideration thus concerns the relationship between the temporal phenomena postulated at the prosodic level and the precise acoustic implementation of these phenomena.

As will be seen later in the chapter, segment or syllable durations and pause phenomena are likely to be simply two sides of the same coin. We shall begin by describing pause phenomena, and in the second part of the chapter, we shall show how pause phenomena can be seen in terms of an integrated theory that relates temporal phenomena to a general, prosodic framework for speech.

Pauses

The Classification of Pauses

From a descriptive point of view, two classifications of pauses are in general use. The first one is a physical and linguistic classification, and the second one is a psychological and psycholinguistic classification.

The Physical and Linguistic Classification

In the traditional linguistic definition, normal speech flow is considered to be interrupted by a *physical pause* whenever a brief silence can be observed in the acoustic signal (i.e., a segment with no significant amplitude). Which exact duration of the silence is considered sufficient for the constitution of a physical pause depends on its linguistic context (for an overview, see in particular Dechert and Raupach, 1980):

— *Intra-segmental pauses* are those which are related to the occlusions of the vocal tract in normal speech production. Example: In the word “happy” (Figure 1), the pause component of the Voice Onset Time (VOT) for the consonant / p / corresponds to a silence of 96 ms.

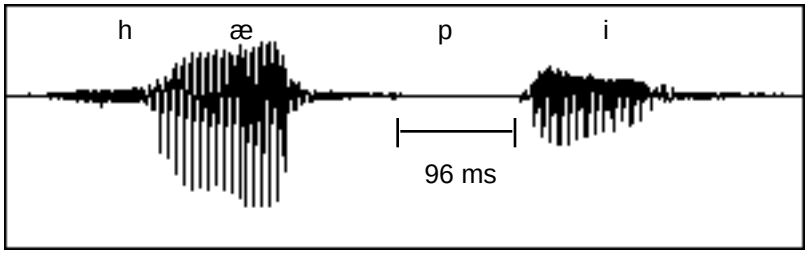

Figure 1. The acoustic signal of "happy" showing an intra–segmental pause.

— *Inter-lexical pauses* are those which may appear between two words. They constitute the first segmentation of speech, or the phrasing that is likely to facilitate the perceptual interpretation of the speech utterance. An example is the differentiation between "a Turkish (carpet salesman)" and "(a Turkish carpet) salesman" (Figure 2).

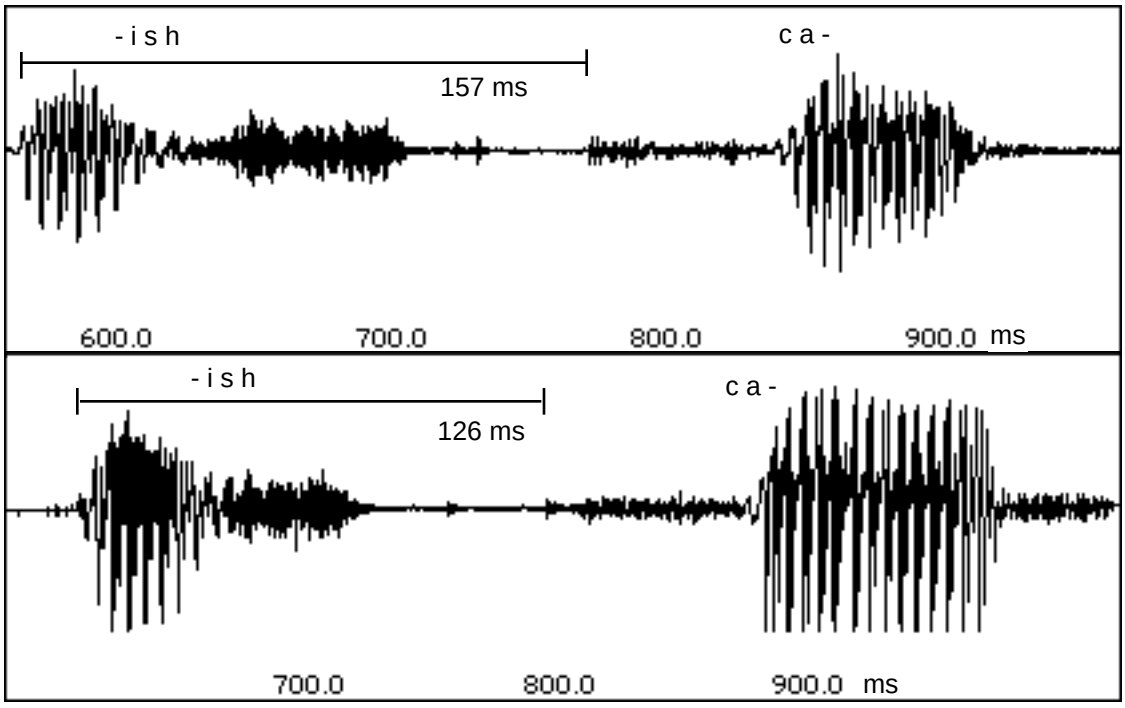

Figure 2:. Acoustic signals corresponding to the pronunciation of "a Turkish (carpet salesman)"(above) and "a (Turkish carpet) salesman" (below).

At the acoustic level, the differentiation of these two sentences is achieved both by variations of the melody (see also Chapter 2, this volume) and by variations of duration. In the first sentence, the duration of the combination of the final syllable duration of the word "Turkish" and the inter-lexical silence equals 157 ms, while in the second sentence, the same acoustic segment measures 126 ms.

It is important to note that what appears to the ear as a simple "pause phenomenon", actually translates at the signal level into a complex manipulation of duration and fundamental frequency that extends over both, the final syllable and the adjoining pause. The

traditional linguistic and psycholinguistic notion of a unidimensional "pause phenomenon" has thus been amended by more recent research. It is nowadays recognised that these so-called "temporal" prosodic phenomena are actually inherently complex, multivariate and supersegmental, since they generally extend over a number of segments or syllables, and involve the interaction between a number of physical parameters.

A Psycholinguistic Classification

Even if the discussion is limited to strictly temporal phenomena, perceived pauses are not really the equivalent of physical pauses. This is due to a law of perception well known to physiologists: Whether in the visual, auditory, or in the tactile domain, the perceptual threshold is situated above the actual physical stimulus. Moreover, amplitude curves measured in detailed perceptual tests (e.g. responses to finely differentiated acoustic;stimuli) differ systematically from curves measured directly on the physical stimulus.

With respect to temporal phenomena, a number of psycho-acousticians have documented the importance of these aspects of perception for our understanding of speech processing (for an overview, see Zwicker and Feldkeller, 1981; Botte *et al.*, 1989). A correct perception of connected speech requires the ability to process about 15 to 30 distinctive sounds ("phonemes") per second. Variations in excess of about 20% of the duration of these sounds have been found to be perceptually relevant to the error-free perception of the speech flow.

Some pauses are more easily perceived than others, and generally, such pauses appear to support particular functions within the message, such as grammatical functions, semantic focus, hesitation, and so on. Also, pauses are more easily perceived if their duration is around 200 - 250 ms. That appears to be the standard auditory threshold for the perception of pauses (Goldman-Eisler, 1968; Grosjean and Deschamps, 1975). In languages where systematic pauses have been observed, two types can be distinguished, so-called "silent" and "filled" pauses:

— *Silent pauses* correspond to the perception of a silent portion in the speech signal. Such pauses may be produced in conjunction with an inspiration, swallowing, any laryngo-phonatory reflex, or a silent expiration.

— *Filled pauses* correspond to the perception of a voiced section in the speech signal. Most filled pauses in such languages as English and French are drawls, repetitions of utterances, words, syllables, sounds, and false starts (Grosjean and Deschamps, 1975; Bloodstein, 1981).

Generally, with normal speakers (speakers with no speech pathology), silent and filled pauses appear *between* words. Pauses of 200 ms and longer are seldom observed within a word.

The Origins of Pauses

Sites and durations of silent and filled pauses are subject to two types of constraints. First, they depend on physiological aspects of speech motor activity, and second, they reflect cognitive processes.

The Temporal Structure of Speech and its Constraints

Like any other human activity, speech production cannot be continuously exercised. Interruptions are necessary, since any specific motor activity runs in parallel and interacts with other types of motor activity. Since motor behaviour is generally acknowledged to be performed according to three successive stages (planning, execution, pausing), any interruption of one activity to admit another usually takes place during the pause segment.

Moreover, physiologically inevitable pauses regularly occur during the inspiration phase of respiration, since phonatory activity is intricately connected with respiratory activity. Finally, speech production may be considered to be a rhythmic activity, where word groups are produced at a particular rate. All these facts likely contribute to the production of regularly spaced pauses.

In addition, a number of other parameters can be shown to influence the occurrence of pauses:

Individual physiological constraints: Since speech motor activity is largely an individual activity, the occurrence of pauses depends to a considerable extent on the specific speaker. In general, weak respiration, low muscular tone, and slow articulatory rate is associated with a greater number of pauses than a rapid articulatory rate and good respiratory capacity.

Temporal constraints and the notion of a rhythmic span: Pauses;tend to occur between rhythmic groups. Dauer shows that the data from a comparative study of syllable-timed and stress-timed languages (e.g. French *vs.* English) support the hypothesis of universal features in rhythmic structure (Dauer, 1983). According to this hypothesis, speech planning is based on a psychological regulatory unit that allows for about two “acts” per second.

Evidence from timing in general motor behaviour tends to support the notion of some type of biological clock monitoring every rhythmic

activity. The durations of some speech segments show two features that suggest the presence of an internal clocking mechanism: first, certain segments are quite regular, and second, the variations for some durations (e.g. the duration of the time between two CV transitions in French) show regular corrections — very much as if constant adjustments are made, so as to approximate certain idealised time frames. According to Barbosa and Bailly (1993), this internal clock is "a time-keeping function used for the synchronisation of impulses transmitted to the muscles". It programs the duration of rhythmic units, and clock regularity is maintained by means of pauses. In other words, the durations of pauses are constrained to a given number of clock units. Obviously, this notion of regularity is very attractive, particularly in the context of a rule-governed natural speech synthesis. However, Keller (1990) recalls that at the biological level, there exists no established link between a neural pulsation of central origin and speech timing. It may thus be safest to interpret the existing evidence as saying that some internal time regulation probably exists, but that its exact physiological mechanism is still unknown.

Dauer furnishes a somewhat similar argument. According to this author, regularity in stress-timed languages is manifested in the stress interval, where the mean interval ranges from 1.9 to 2.3 stresses per second. These stresses occur at much greater regularity than do syllables, where the mean inter-syllable interval ranges from 4.5 to 8 syllables per second. In Dauer's view, rhythmic grouping applies to all languages and interacts with their syllable structure, their rules of vowel reduction and accentuation. Based on such notions of durational regulation, it can be proposed that rhythmic grouping of speech favours the occurrence of pauses at certain intervals.

The Function of Perceived Inter-lexical Pauses (in Excess of 200 ms)

Beyond the largely physiological origins of the fairly regularly occurring pauses cited above, it is also possible to identify a number of cognitive origins for pauses. Although these latter pauses are less regular in occurrence, they do show patterning and could thus be used in generating natural-sounding synthetic speech.

Pauses as a reflection of cognitive activity: According to Goldman-Eisler (1968, 1972), a pause is the external reflection of some of the cognitive processes involved in speech production. In this sense, pauses provide additional time during which the final output can be planned and programmed. This hypothesis explains some common observations, such as when a speaker thinks a long time before providing a very quick, clear, and well-constructed reply. On the other hand, it can also be observed that sometimes, a speaker begins to reply at once, and then has to stop or retrace his steps to clarify his message. In this case, the hypothesis

proposes that "speech has raced ahead of cognitive activity" and that the pause reflects the time needed for the cognitive planning process to catch up.

The Goldman-Eisler hypothesis further predicts commonly observed differences between spontaneous and read speech, in that spontaneous speech is much more conducive to pauses of cognitive origin than is read speech. In this sense, these are not simply nuisance variables. Since they probably reflect cognitive activity, they contribute to the hearer's understanding of how the speaker is structuring the utterance. And since pauses apparently participate in rendering human communication more intelligible, it can be postulated that some future, advanced automatic speech recognition mechanisms could conceivably benefit from this type of information to improve recognition performance.

Pauses acting as "beacons" for utterances: A number of psycholinguistic investigations have furnished some clear indications on how pauses are patterned. In a study on French, for example, Grosjean and Deschamps (1975) have shown that the more complex the communicative task, the greater the number of pauses (see Table 1). When describing cartoons, pauses are both longer and more frequent than when responding in an interview (Grosjean and Deschamps, 1975).

Table 1. Duration of pauses according to linguistic task.

	Mean duration of silent pauses:
Description of a cartoon:	1320 ms
Interview:	520 ms

Moreover, in both communicative situations, the duration of these pauses is quite impressive (1320 and 520 ms), particularly when considered in relation to the duration of a typical syllable (some 200-300 ms), or to that of a typical vowel (some 100-150 ms) (O'Shaughnessy, 1981). In other words, pauses "stick out like sore thumbs", and thus may occupy "beacon" positions in speech, serving to structure the entire utterance for both speaker and listener. By subdividing speech into smaller segments, pauses probably contribute a great deal to the improvement of speech comprehension.

In addition, there exists a relationship between the duration and frequency of pauses on the one hand, and the syntactic constituents hierarchy on the other. In a reading-aloud task, for example, there is a tendency for a pause to be longer, the more "profound" the syntactic boundary. Still, it will be seen below that temporal segmentation is not really equivalent to the syntactic structure of utterances.

Situational constraints: Finally, a consideration of the situational context is also important, because the temporal pressure on the speaker can favour or hinder his expressive capacities. (For example, it is difficult

to furnish an important information very quickly in a noisy room.) The more difficult the communicative situation, the more pauses, hesitations, and stuttering events are likely to occur.

Speech production being a rhythmic activity as well as the reflection of the underlying cognitive processes, speakers thus produce pauses spontaneously. This and the links between temporal and syntactic structures lead us believe that perceived inter-lexical pauses should be predictable and open to algorithmic implementation in speech synthesis.

Pauses and the Notion of Verbal Fluency

The phenomena described here can be captured by the overall notion of *verbal fluency:* a speaker is *fluent* when he speaks easily, with smooth onsets and transitions, and at a relatively rapid clip (Pfauwadel, 1986). Conversely, a speaker is to some degree *dysfluent*, if he is hesitant, produces pauses at inappropriate places and makes speech errors. It is a well-documented fact of human communicative behaviour that speakers can range from extremely fluent to extremely dysfluent. Even without being affected by any manifest neurological impairment, some speakers are barely comprehensible because of excessive hesitations, pauses and speech errors.

From Extremely Dysfluent to Extremely Fluent

It is useful to think of speakers' fluency as if it were distributed on a Gaussian curve. Extremely dysfluent speakers would be located at the left of the curve, followed by various degrees of hesitant speakers, passing through the majority of speakers who are reasonably fluent, and ending at the right in a small group of extremely fluent speakers.

In spontaneous speech, it is not unusual to hear someone saying one word instead of another, or mixing up the syllables of a word. These are so-called *performance errors* or *normal dysfluencies.* In contrast with pauses which are quite regular, performance errors are unique. They are produced at specific moments, in particular situations, and they will not generally be produced again in other conditions. However, the irregular occurrence, as well as the frequent correction of this type of error leaves no doubt that the speaker really knew what he intended to say. In contrast to errors produced by children or foreign language learners, his "speech competence" is not in question.

The speech impairments at the left of the curve also include various degrees of stuttering. In these cases, the temporal structure of speech diverges from the expected pattern: the frequency, the durations and the sites of pauses are abnormal. When a *non-stuttering* speaker hesitates, his

(silent/filled) pauses are generally located at syntactic or prosodic boundaries (Zellner-Bechel, 1992). A near-pathological or *pathological stutterer*, on the other hand, produces dysfluencies that are often located far from syntactic or prosodic boundaries (Starkweather, 1987). An abnormal segmentation tends to perturb the listener, sometimes quite strongly, while a normal segmentation is hardly ever perceived consciously. This permits the extrapolation that *abnormal pause insertion* can be as destructive to the perceptual decoding of an utterance, just as a normal use of pauses can be useful to its understanding.

Even highly fluent speakers vary the temporal structure of their speech. Miller *et al.* have shown that in spontaneous speech, a speaker can vary his rate considerably within the same utterance, whereby the more extensive rate variations are implemented by manipulations of number and duration of pauses (Miller *et al.*, 1984). Furthermore, rate variations are related to the use of variants of speech sounds, since at a fast speech rate, sounds (or "allophones") diverge more from "ideal" (non-reduced) phonetic variants of a given phoneme than at a slow speech rate (Lacheret-Dujour, 1991).

This characterisation of fluency in the speech flow poses a distinct challenge to speech synthesis and speech recognition. Neither technology currently exploits the occurrence of pauses and speech errors. In the future, it may well be of interest to generate patterned (normal) pauses in synthetic speech. The notion of "fluency" is directly interrelated with the prediction of quite a number of phonetic parameters, such as speech rate, pauses, errors and vowel reduction. If all these parameters are some day controlled in a coherent manner in speech synthesis, the artificial voice will probably convey an impression of greater ease and fluency of speech. Similarly, automatic speech recognition devices developed to deal with spontaneous speech will inevitably have to learn to deal with the phenomenon of pauses, speech errors and speech error repair some day.

To prepare the terrain, it may be useful to examine some details of the temporal structure of speech, as well as various predictive algorithms. This is what we turn to next.

The Durational Structure of Speech

We saw that speech rate constitutes an essential aspect of verbal fluency. Speech rate is determined not only by pauses, but also by the rate of articulation (Grosjean and Deschamps, 1975). Variations in the rate of articulation are induced by several factors, primarily by variations of the durational structure found in speech.

There is an extensive literature on this question, but we shall only refer to the well-known study of Klatt (1976), since his study was oriented

towards the support of speech synthesis. In this study, the factors that had been found to influence the temporal structure in spoken English sentences were inventoried in the literature. Each factor was evaluated for its capacity to provide perceptual cues sufficient to make linguistic decisions. Seven factors were retained (and subsequently integrated in the statistical model of durational control used in MITalk):

1. Extralinguistic factors (e.g. speaker mood)
2. Discourse factors (position within a paragraph)
3. Semantic factors (emphasis and semantic novelty)
4. Syntactic factors (phrase structure lengthening)
5. Word factors (word-final lengthening)
6. Phonological / phonetic factors (inherent phonological duration, stress)
7. Physiological factors (inherent duration and incompressibility)

It can be seen that duration influences every level of speech production. However, it is useful to distinguish between two levels of durational control, extrinsic and intrinsic (Ferreira, 1993). Units of word length (lexemes, inflected words, fixed expressions, etc.) are said to have a set of *intrinsic* durations which are presumably stored in a mental lexicon. Each time they are used, the basic *distribution* of duration of its various segments will be roughly the same. As these units are integrated into larger entities (phrases, utterances), they get "stretched" and "squeezed" in accordance with the requirements of larger speech demands[1]. These larger demands correspond to an *extrinsic* level of durational control. There will be more or less expansion, depending on where the word occurs in the utterance, on whether the word is emphasised or not, and on what grammatical group the word belongs to.

Therefore, the first task in a text to speech system is to parse the sentential structure into "natural" word groups. Pauses tend to occur between such word groups. The correct prediction of sites and extent of silent and filled pauses is largely determined by the extent of prosodic units, the so-called *performance structures.*

A Parsing Tool: The Performance Structures

A performance structure is a psycholinguistic structure that captures the various degrees of cohesion between the words of an utterance. For

[1] It is to be noted that the "streching" and "squeezing" does not apply to all segments equally. Stop consonants, for example, are much less subject to temporal modification than other types of segments, such as vowels and fricative consonants (Fujimura, 1981).

example, in the preceding sentence, there is much greater cohesion between the words “the”, “various” and “degrees” than between “structure” and the succeeding word “that”. By “cohesion” is meant frequency of co-occurrence, semantico-syntactic relationships (such as determiner-noun or adjective-noun relationships), and syntactic relationship (like singular / plural agreement). 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 an utterance. Essentially convergent performance structures have been demonstrated by several types of psycholinguistic experiments, such as memory tasks or intuitive parsing tasks performed by non-linguists, as well as by the measurement of pauses and syllable durations in speech. Since the empirical measurement of pauses and syllable durations are of the most direct relevance in the present context, we shall concentrate on this type of evidence.

Starting with Grosjean *et al.* (1979), a number of authors have shown that occurrence and lengths of pauses are strongly correlated with the degree of inter-lexical cohesion, that is, pauses tend to be long and frequent between words that show relatively little cohesion and they are much shorter and less frequent between words that are strongly interdependent. This grouping of words appears to be independent of respiratory constraints: The same structures are found in a reading aloud task of sentences produced with and without a respiratory break. Besides, whatever the experimental method, a similar sentence segmentation is obtained.

Table 2. Word grouping.

The verbal stream seen in the speech signal
speech flow / **pause** / speech flow / **pause** / speech flow / etc.
corresponds statistically, semantically and syntactically to
words with strong cohesion / **pause** / words with strong cohesion / **pause** / etc.
and corresponds psycholinguistically to
performance structure / **pause** / performance structure / **pause** / performance structure / etc.

To be able to use this notion in text-to-speech synthesis, reliable predictors must be identified. In this process, the main problem consists of identifying *automatically* where a performance structure begins and ends.

According to Grosjean (Grosjean and Dommergues, 1983; Monnin and Grosjean, 1993), performance structures have three main characteristics:

(1) *"Eurhythmy":* the basic units tend to be of the same length.

(2) *Hierarchy:* the basic units are enclosed into larger units that are themselves incorporated into even larger units.

(3) *Symmetry:* performance structures tend to be balanced, and the major pause is located around the middle of the sentence.

A second problem concerns the internal organisation of this type of structure. To answer this question, it is necessary to clarify how the postulated psycholinguistic performance structures (and the subsumed temporal speech structures) relate to linguistically-based syntactic structures and / or to phonetically-based prosodic structures.

Are Performance Structures Based on Syntactic and / or on Prosodic Structures?

Performance Structures and Syntactic Structures

Given the extensive work performed on syntactic structures during the 1960s and 1970s, it was natural that performance structures were initially assimilated to syntactic structures. Since performance structures seem to be organised in terms of a cohesion between various words, this cohesion was seen as directly related to that expressed by syntactic structures, structures that are based on specific criteria for capturing grammatical inter-relations between lexical units (Martin, 1980).

For this reason, a number of initial studies were directed at generating performance structures from a syntactic analysis of the sentence (e.g. Grosjean, 1980a; Grosjean and Dommergues, 1983). It is recalled that syntactic structures are generally seen as upside-down trees, where the lower nodes (branching points) connect strongly related, typically proximal elements and higher nodes connect larger, less directly related groups of syntactic units. Within this type of structure, transitions between portions of the tree depending on different higher nodes were associated with longer silent pauses, according to the principle of "the more profound the transition, the longer the pause". However, obtained results were unsatisfactory, because syntactic and psycholinguistic structures were not found to be homomorphic (Grosjean and Dommergues, 1983). Mismatches between syntactically proposed structures and empirically derived performance structures were frequent.

This result was independent of the exact syntactic theory that was assumed.

More specifically, it was found that syntactic structures are insensitive to the length of their constituents, while the main performance units tend to be approximately equilibrated, in the sense that they contain about the same number of words, and that the lengths of their units are similar. If a sentence has, for example, a *short noun phrase* followed by a *long verb phrase*, the major syntactical boundary is situated between the two phrases. However according to the psycholinguistic evidence, the major "performance" break is located between the verb and the rest of the verb phrase. On the other hand, a *long noun phrase* and a *short verb phrase* are subdivided differently, so that the constituents end up being more or less balanced.

For example, when French speakers are asked to segment a sentence into intonational groups or units, they tend to balance out the number of syllables in each group (Grégoire, 1899; Grosjean and Dommergues, 1983; Monnin and Grosjean, 1993; Fónagy, 1992; Martin, 1992). The mean length of these units, in a reading aloud task of simple sentences, was estimated to be around 3.46 syllables with a standard deviation of 1.43 (Monnin and Grosjean, 1993).

The syntactic complexity analysis of a sentence is no doubt an important factor, but in itself, it has not turned out to be a sufficient predictor of performance structures. Lehiste (1972) examined the effects of morphological and syntactic boundaries on the temporal structure of spoken utterances. She concluded that temporal readjustment processes tend to ignore morpheme and word boundaries. The durational structure is, according to Lehiste, conditioned by the number of syllables, rather than by either the number of segments, or by the presence of boundaries. Moreover, Ferreira (1993), has shown by means of an elegant experiment that the temporal structure of a sentence is closely related to its prosodic structure. If a syntactic variation (number of syntactic boundaries following a key word) is introduced into the given textual material, temporal variables are not significantly affected. However, if a prosodic variation (a change of emphasis) is introduced, temporal variables such as final-word durations and pauses are notably affected. We may conclude that performance structures depend both on linguistic variables (such as syntactic boundaries), and on psycholinguistic variables (such as length or symmetry).

Performance Structure and Prosodic Units

If prosodic factors are related to, but not homomorphic with syntactic structure (for an overview, see Hirst *et al.*, in press), and if performance structures are related to, but are not veritably congruent with syntactic

organisation, it remains to be seen if a closer relationship can be established between performance structures and prosodic units. In terms of our schema, that would give:

Table 3. Performance structures and prosodic units.

performance structure / **pause** / performance structure / **pause** / performance structure / etc.
as equivalent to
prosodic unit / **pause** / prosodic unit / **pause** / prosodic unit / etc.

There are some indications that this relationship holds. According to Hirst (in press), silent and filled pauses constitute the acoustic;marks that enclose the prosodic units (or as in the literature: “prosodic phrase”, “prosodic structure”, etc.). As the suprasegmental and audible tissue of speech, such prosodic units can be defined by the *relations* existing between prosodic parameters of speech (stress, intonation, duration, etc.). Some similar results appear to come out of our own algorithmic work (see below). In view of the apparent link with prosody, it may thus be useful to review at how humans appear to process prosodic structures.

How are Prosodic Structures Generated in the Human Mind?

The main question concerns the manner in which a performance structure – *i.e.*, the phrasing of the sentence – is constructed as an utterance is prepared for output. According to Bachenko and Fitzpatrick (1990), there exists a *neutral phrasing*, that is, a sentence-level phrasing pattern that is independent of discourse semantics. According to these authors, it is upon this neutral pattern that all the variants like focalisation, emphasis, etc., are grafted. They suggest that this discourse-neutral phrasing is built upon a knowledge of syntactic constituents, of string adjacency, and of phrase length.

In this way, recent phonological theories conceive of the sentence as a set of prosodic units that combine phonological and syntactic units. Each prosodic unit specifies rhythmic characteristics at the syllabic level as well as at the word-, the phrase-, and the sentence levels. For example from a prosodic point of view, the word is a set of feet, *i.e.*, a strong-to-weak syllable pattern. Such words are grouped into a *phonological phrase* (Selkirk, 1984). This phonological phrase is not necessarily isomorphic with the syntactic phrase. At the end, phonological phrases group into intonational phrases, which in turn combine to make an utterance.

The question is at which point during the process of speech production this neutral phonological phrase is encoded. Ferreira (1993) considers that the prosodic constituent structure is generated *without knowledge* of the segmental content, just as in Levelt's model of language production (for further details, see Levelt, 1989). According to this hypothesis, abstract timing intervals are assigned to various sentence locations early on in the speech planning process. Later, segmental factors are taken into account, but word and pause durations are in a trade-off relationship, so as to respect the total duration of the allocated interval. At the phonetic level, a long word takes up more space in the timing interval than a short word, but a greater word length also implies a shorter subsequent pause. That means that within an interval, word and pause lengths are inversely related. The size of a timing interval is determined by the number of prosodic constituents. The more prosodic constituent boundaries that end on a given word, the longer the timing interval for that word will be.

A general agreement concerning temporal phenomena can thus be noted. It can now be stated with some certainty that phrasing is related to prosodic structures, and that rhythmic patterns permit to establish a connection between temporal phenomena and the segmental content. That is the principal foundation upon which speech synthesis algorithms can be built.

The Automatic Generation of Durations: Algorithms

Once again, the reader will not find a complete overview of this area here. The following algorithms are, from our point of view, simply the most important attempts to generate durations automatically on the basis of text input. Two types of duration models can be distinguished: *rule-governed systems* and *statistical systems.*

Rule-Governed Systems

In these approaches, a number of rules — generally sequential rules — are applied to linguistic material which is characterised by an intrinsic duration. It results in an output duration. These output durations are supposed to take into account various phenomena such as linguistic constraints (like syntactic factors) or psycholinguistic constraints (like length).

The following algorithms (GGL, CPC, PHI, Bachenko and Fitzpatrick, and MG) are all built on the same basic notions, as they use a number of concepts from generative syntax.

Grosjean Grosjean and Lane (1979) — GGL. The aim of this algorithm was to predict performance structures. For each word boundary, this cyclic algorithm matches an index of linguistic complexity to a measure of the distance to the midpoint of the segment.

The mean correlation of 0.83 between predictions and observations leads to the conclusion that this simple algorithm is a relatively good predictor of durations, even of the pause data. However, the syntactic analysis tends to overestimate the importance of inter-lexical function words (Gee and Grosjean, 1983).

Cooper and Paccia Cooper (1980) — CPC. Their algorithm contains 14 rules to predict, for English, the probability of a pause occurring at each inter-lexical boundary, as well as segmental lengthening, pause and segment durations. It essentially accounts for the depth of syntactic nodes.

The mean correlation obtained between measures and predictions is about 0.75. When it is applied to French, the correlation decreases to 0.57. The main objections are that the rule application is not clear enough, and that there is no bisection compounds for mid–sized sentences (Gee and Grosjean, 1983).

Gee and Grosjean(1983) Phonological Phrase Algorithm — PHI. This algorithm suggests that two factors govern the temporal structure of the sentence: The syntactic structure and the information content-based distinction between function and lexical words. The function words are considered to be weak syllables. This algorithm does not require the whole tree structure of a sentence to start segmentation into prosodic units.

The mean correlation is reported to be very high (0.96), probably because this algorithm integrates phonological, syntactical and prosodic information. It demonstrates that performance structures reflect prosodic structures. When it is applied to French, the major prosodic boundaries are rather well predicted, except that the post–verbal boundary is overestimated.

Bachenko and Fitzpatrick (1990) — BF. Influenced by the former algorithm, these authors conceived a system for English capable of predicting phrasing by a process of localising phrase boundaries. First, phonological words and phrases are identified. Then, boundary salience rules are applied, assigning a relative strength — i.e., perceptibility — to each phrase boundary according to syntactic labelling, length, and adjacency.

Their results indicate that 80% of the primary boundaries — i.e., the most easily perceived boundaries — were correctly predicted. The system's limitation concerns its parsing errors. For example, incorrect part of speech assignments and incorrect analyses of pre–head modifiers have been identified.

Monnin and Grosjean (1993) — MG. This algorithm specifically developed for French is also promising, because the mean correlation

between the performance structures of nine read sentences and the prediction of these structures with the algorithm is about 0.94. Major and minor prosodic boundaries are correctly predicted. Consequently, the sentence is segmented into prosodic units that are not necessarily equivalent to surface syntactic structures. However, the algorithm has not been extensively tested on other sentences.

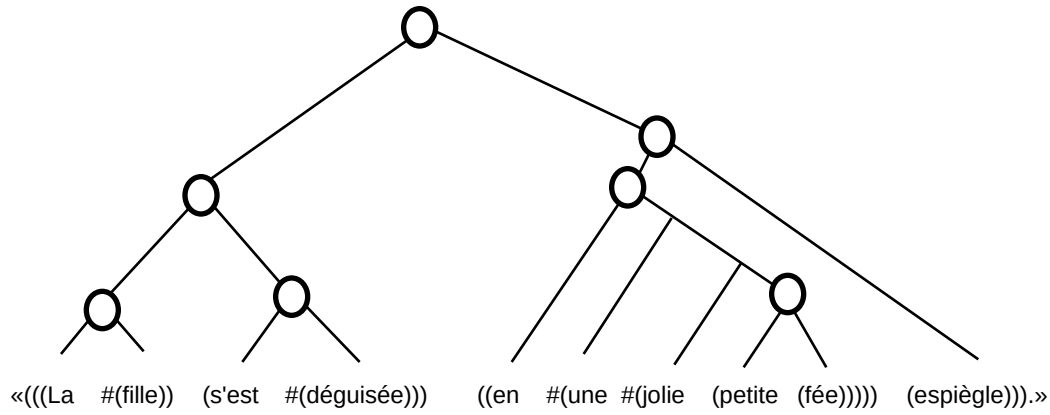

Figure 3. A performance structure tree according to the Monnin and Grosjean rules. A single node separates «petite» and «fée», which predicts a short final syllable and adjoining pause for «petite». By contrast, five nodes separate «déguisée» and «en», which predicts a much larger final syllable and pause duration at this major juncture. The #-mark specifies an attachment of grammatical to lexical words. Performance structure trees are quite different from syntactic structure trees, no matter the theoretical tradition.

The Keller-Zellner (KZ) Rules (1993): Initially derived from the MG algorithm and conceived for French, this algorithm is quite different in form as well as in content. Basically, the aims of these rules are to satisfy the criteria of *simplicity*, respect of *psycholinguistic plausibility*, and *high predictive capacity* for the data sets at hand.

Prosodic constituents are formed on the basis of simple proximal syntax. No syntactic structures more complex than those applying to a single phrase are required. Prosodic groups can be identified by the application of steps 1 and 2 of the Monnin-Grosjean rules.

Final syllable+pause durations increase in duration as the constituent proceeds. The increase proceeds from an empirical minimum to an empirical maximum. The initial hypothesis calls for equal steps. Increased durations correspond to a slowing down, which is a commonly observed phenomenon in speech.

Rhythmic alternance was observed for two locations: post-verbally and in the middle of 4-6 word constituents. Rhythmic alternance occurs when one element is lengthened more than strictly required. As a consequence, the following element must be shortened “in order to conclude the constituent in time”. Concretely, this amounts to postulating an *inversion of durations* for the word pair involved in the alternance.

The resulting algorithm is quite simple and is reproduced at the end of this chapter. Correlations with the Caelen-Haumont and the Monnin-Grosjean data sets are reported in Table 4. It is found that correlations are quite regular. They never dip below a linear correlation of.7, and generally tend to be found in the 0.8 range.

Table 4. Linear correlations between predicted and measured final syllable+pause durations according to two sets of rules.

	Monnin-Grosjean		**Keller-Zellner**	
Caelen-Haumont Data Set	**Normal**	**Slow**	**Normal**	**Slow**
Sentence 1	.786	.895	.862	.845
Sentence 2	.289	.375	.811	.829
Sentence 3	.925	.808	.878	.751
Mean	**.667**	**.693**	**.850**	**.808**
Monnin-Grosjean Data Set	**Normal**	**Slow**	**Normal**	**Slow**
Sentence 1	.890	.674	.873	.835
Sentence 2	.914	.796	.886	.954
Sentence 3	.981	.886	.773	.892
Sentence 4	.961	.826	.798	.850
Sentence 5	.947	.736	.827	.872
Sentence 6	.984	.711	.812	.835
Sentence 7	.931	.841	.754	.906
Sentence 8	.940	.585	.870	.809
Sentence 9	.968	.808	.701	.818
Mean	**.946**	**.763**	**.810**	**.863**

An inspection of the evolution of F_0 and energy values at the end of prosodic constituents postulated here shows some regularities. F_0 values rise at the end of each constituent, except for the sentence-final constituent. Energy values fall regularly at the end of each constituent. This suggests that the temporal structure characterized here interacts directly with control over F_0 and energy.

As they cannot predict the *whole* temporal structure of an utterance, the previous models are obviously incomplete. Even if inter-lexical pauses and word-final lengthenings are generated satisfactorily, these algorithms neglect syllabic and segmental durations.

Statistical Systems

Within the perspective of text-to-speech systems, it is important to control speech timing at each level of the sentence generation process. The MITalk system — the English text-to-speech system developed at the Massachusetts Institute of Technology (Allen *et al.*, 1987) — satisfies this

requirement and thus provides an excellent point of departure. Although this system belongs to an older generation of synthesizers and sounds quite artificial by today's standards, its temporal structure is based on an influential approach first developed by Klatt (1976). It is a statistical model built around segmental durations, *i.e.*, durations for individual phonemes.

On the basis of data collected in a variety of projects conducted by several authors, Klatt's vowel model begins by calculating an inherent phonological duration for a given segment. This duration is then shortened or lengthened as a function of the succeeding segment, its position within the phrase, the presence or absence of stress, and word length. Lehiste's non-sense word data (cf., Klatt, 1976) was re-analysed, and it was found that just four rules in Klatt's model can account for 97% of the total variance of the measured vowel durations.

A somewhat similar model was proposed by O'Shaughnessy (1981, 1984). This is probably the most important statistical model for spoken French text. On the basis of numerous readings of a short text containing all phonemes of French, a model of durations of acoustic; segments suitable for synthesis by rule was proposed. In this model, 33 rules specify basic durations for various classes of segments, as well as modifications to this basic duration as a function of phonetic context.

For sound classes that do not involve prepausal lengthening, the model was able to predict the durations for 281 segments of a text, with a standard deviation of 9 ms. But it was less accurate for the prediction of prepausal vowel durations, because of the greater variability of these segments in such positions. Moreover, this model was not able to predict silent inter-lexical pauses.

These two statistical models are constructed around the same essential hypothesis, an hypothesis which is open to an important critique. The authors assume that speech timing phenomena can be captured by the segment, as if this unit "possesses an inherent target value in terms of articulation or acoustic;manifestation" (Fujimura, 1981). However, recent measures have indicated that syllable-sized durations are generally less variable than subsyllabic durations, and thus represent more reliable anchor points for the calculation of subsyllabic durations (Barbosa and Bailly, 1993). This approach to duration receives further support from the observation that stress variations and variations of speech rate tend to modify at least syllable-sized units, certainly more than single segments. The durations of coarticulation phenomena could also profit from this reorientation of perspective. It may be more profitable to see segments and their coarticulatory phenomena from the perspective of integrating a number of segments into a *syllable*, than from that of the atomic durational measure represented by the phoneme.

Independent of whether the calculation begins with the segment or with the syllable, a statistical approach to duration is very promising.

Results of this type suggest that a likely next step in this area of research is the development of a robust statistical model, capable of predicting the entire durational structure of a sentence.

Conclusion

It is estimated that speech synthesis will sound more fluent, will be more pleasant to listen to, and will likely be more intelligible when silent and filled pauses are systematically integrated into the verbal stream. Since these pauses enclose prosodic units — which we have equated with performance structures — pauses can be predicted by the same mechanisms that let us predict performance structures. Algorithms have been developed that test these predictions, and which are presently being tested with respect to various types of speech material.

These advances are part of a larger effort to develop models capable of controlling various aspects of speech fluency in TTS systems. Issues in pausing and in speech error repair are also of importance to automatic speech recognition, since future systems for the understanding of spontaneous speech will have to show sufficient intelligence to deal with speech repair, and may well profit from the regularity of the pause and syllable duration patterns found in the temporal structure of speech.

The Keller-Zellner Algorithm

(1) Identification, from left to right, of the *nuclei* of the prosodic constituents: nouns, verbs and free-standing adjectives, adverbs and pronouns (such as “La chemise est *sale*”, “c’est *bien*”, “pense à *ça*”).

(2) Creation of the *prosodic constituents* by grouping the words around the nucleus. All words to the left of the nucleus are attached to the right-lying nucleus, except for post-posed adjectives and post-posed pronouns which are attached to the left-lying nucleus (“la chemise *blanche*”, “donne-*lui*”).

(3) Calculation of *predictions for final syllable+pause durations*. Within each prosodic constituent, durations increase from a minimum to a maximum duration. Initially, the increase is assumed to occur in equal steps. (The minimum and maximum are assumed to be 50 and 350 ms in normal speech, 50 and 525 ms in slow speech.) The first final syllable in a constituent has a duration of minimum+step size ms.

(4) *Rhythmic tradeoffs:*

1. *Post-verbal trade-off:* When a constituent follows a verb and there are at least two words prior to the nucleus, the final syllable duration of the first word is lengthened with respect to that of the second word. (Exchange durations for words 1 and 2.)

2a. *Rhythmic alternance:* If a constituent is four or more words long, and if word 3 is two or more syllables long, word 2 is lengthened with respect to word 3. (Exchange durations for words 2 and 3.)

2b. *If rule 1 has already applied:* If a constituent is four or more words long, and if word 4 is two or more syllables long, word 3 is lengthened with respect to word 4. (Exchange durations for words 3 and 4.)

3. *Single-word constituents:* Constituents containing a single word show reduced final syllable durations. (Reduce durations for single word constituents by 50 ms.)

(5) *Measure of final syllable+pause.* The measure begins with the vowel of the final syllable and ends at the end of the pause. It includes whatever intervening consonant may occur, but it excludes the characteristic optional schwa of French méridional speakers (as in «biologist**e**»). Excluding the optional schwa permitted us to make direct comparisons of data sets from northern and méridional speakers. Resulting time measures were very similar. For a limited data set, the intervening consonant was suppressed. However, resulting durations were found to show greater variability than those that included the consonant. Measures for sentence-final words were only known for a few sentences and were thus set to 0 in all cases for statistical purposes.

References

Allen, G.D. (1975). Speech rhythm: Its relation to performance universals and articulatory timing. *Journal of Phonetics, 3,* 75-86.

Allen, J., Hunnicutt, M.S., & Klatt, D. (1987). *From text to speech. The MITalk system.* Cambridge, England: Cambridge University Press.

Bachenko J., & Fitzpatrick E. (1990). A computationnal grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics, 16,* 155-170.

Barbosa, P., & Bailly, G. (1993). Generation and evaluation of rhythmic patterns for text-to-speech synthesis. *Proceedings of an ESCA Workshop on Prosody* (pp. 66-69). *Lund, Sweden.*

Bloodstein, O. (1981). *Handbook on stuttering.* Chicago: National Easter Seal Society for Crippled Children and Adults.

Boomer, D.S., & Ditmann, A.T. (1962). Hesitation pauses and juncture pauses in speech. *Language and Speech, 5,* 215.

Botte, M.C., Canévet, G., Demany, L., & Sorin, C. (1989). *Psychoacoustique et perception auditive.* Série audition. Paris: Inserm / Sfa / CNET.

Butcher, A. (1980). Pause and syntactic structure. In W. Dechert & M. Raupach (Eds.), *Temporal variables in speech* (pp. 86-90). Mouton.

Caelen-Haumont, G. (1991). *Stratégies des locuteurs et consignes de lecture d'un texte: analyse des interactions entre modèles syntaxiques, sémantiques, pragmatique et paramètres prosodiques.* Thèse d'Etat, Aix-en-Provence.

Chafe, W. (1980). Some reasons for hesitating. In W. Dechert & M. Raupach, (Eds.), *Temporal variables in speech* (pp. 169-180). Mouton.

Cook, M., Smith, J., & Lalljee, M. (1974). Filled pauses and syntactic complexity. *Language and Speech, 17,* 11-16.

Cooper,W., & Paccia Cooper, J. (1980). *Syntax and Speech.* Cambridge, MA: Harvard University Press.

Dauer, R.M. (1983). Stress-timing and syllable timing reanalyzed. *Journal of Phonetics, 11,* 51-62.

Dechert, W., & Raupach, M. (Eds.), *Temporal variables in speech.* Mouton.

Fraisse, P. (1974). *La psychologie du rythme.* PUF, Paris.

Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological Review, 2,* 233-253.

Fónagy, I. (1992). Fonctions de la durée vocalique. In P. Martin (Ed.), *Mélanges Léon* (pp. 141-164). Editions Mélodie-Toronto.

Fujimura, O. (1981). Temporal organisation of articulatory movements as a multidimensional phrasal structure. *Phonetica, 38,* 66-83.

Gee, J.P., & Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology, 15,* 411-458.

Goldman–Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech.* New York: Academic Press.

Goldman–Eisler, F. (1972). Pauses, clauses, sentences. *Language and Speech, 15,* 103-113.

Grégoire, A. (1899). Variation de la durée de la syllabe en français. *La Parole, 1,* 161-176.

Grosjean, F. (1980a). Linguistic structures and performance structures: Studies in pause distribution. In W. Dechert & M. Raupach (Eds.), *Temporal variables in speech* (pp. 91-106). Mouton.

Grosjean, F. (1980b). Comparative studies of temporal variables in spoken and sign languages: A short review. In W. Dechert & M. Raupach (Eds.), *Temporal variables in speech* (pp. 307-312). Mouton.

Grosjean, F., & Deschamps, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français. *Phonetica, 31,* 144-184.

Grosjean, F., & Dommergues, J.Y. (1983). Les structures de performance en psycholinguistique. *L'Année psychologique, 83,* 513-536.

Grosjean, F., Grosjean, L., & Lane, H. (1979). The patterns of silence: Performance structures in sentence production. *Cognitive Psychology, 11,* 58-81.

Hirst, D., & Di Cristo, A. (in press). *Intonation systems: A survey of twenty languages.*

Klatt, D. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59,* 1208–1221.

Keller, E. (1990). Speech motor timing. In W.J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 343-364). Kluwer Academic Publishers.

Keller, E., Zellner, B., Werner, S., & Blanchoud, N. (1993). The prediction of prosodic timing: Rules for final syllable lenthening in French. *Proceedings ESCA Workshop on Prosody* (pp. 212-215). *September 27-29. Lund, Sweden.*

Lacheret–Dujour, A. (1991). Le débit de la parole: un filtre utilisé pour la génération des variantes de prononciation en français parisien. *Actes du XIIème Congrès International des Sciences Phonétiques* (pp. 194-197). *Aix en Provence.*

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America, 51,* 2018–2024.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Martin, Ph. (1980). L'intonation est-elle une structure congruente à la syntaxe? In M. Rossi et al. (Ed.), *L'intonation: de l'acoustique à la sémantique* (pp. 234-271). Klincksieck.

Martin, Ph. (1992). Il était deux fois l'intonation. In P. Martin (Ed.), *Mélanges Léon* (pp. 293-304). Editions Mélodie-Toronto.

Miller, J., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica, 41,* 215-225.

Monnin, P, & Grosjean, F. (1993). Les structures de performance en français: caractérisation et prédiction. *L'Année Psychologique, 93,* 9-30.

O'Shaughnessy, D. (1981). A study of French vowel and consonant durations. *Journal of Phonetics, 9,* 385-406.

O'Shaughnessy, D. (1984). A multispeaker analysis of durations in read French paragraphs. *Journal of the Acoustical Society of America. 76,* 1664-1672.

Pfauwadel, M.-C. (1986). *Etre bègue.* Paris: Retz.

Selkirk, E.O. (1984). *Phonology and syntax: The relation between sound and structure.* MIT Press, Cambridge, MA.

Starkweather, W.C. (1987). *Fluency and stuttering.* Prentice Hall.

Zellner-Bechel, B. (1992). Le bé bégayage et euh... l'hésitation en français spontané. *Actes des 19eme J.E.P.* (pp. 481-487). Bruxelles.

Zwicker, E., & Feldkeller, R.(1981). *Psychoacoustique: l'oreille récepteur d'information.* Collection technique et scientifique des télécommuications. Paris: Masson.

INDEX