Zellner Keller, B. & Keller, E. (Forthcoming). Representing Speech Rhythm. *Working Papers of COST 258.* Université de Lausanne.

REPRESENTING SPEECH RHYTHM

B. Zellner Keller, and E. Keller

LAIP, IMM, University of Lausanne,
1015 Lausanne, Switzerland
Phone: +41 21 692 30 23 Fax: +41 21 692 30 45
Brigitte.ZellnerKeller@imm.unil.ch, Eric.Keller@imm.unil.ch

KEYWORDS

Rhythm, Prosody, Speech Synthesis, Speech temporal organisation

ABSTRACT

The issue of representing speech rhythm is understood in this paper as the search for relevant primary parameters that will allow the formalisation of speech rhythm. Current speech synthesisers show that phonological models are not satisfactory with respect to the modelling of speech rhythmicity. Our analysis indicates that this may be in part related to the formalisation of rhythmic representation. Based on the observation of other communicative systems facing the problem of representing rhythm, parameters are described for representing speech rhythmic structures.

INTRODUCTION

Speech rhythm usually designates a complex physical and perceptual parameter. It involves the coordination of various levels of speech production (e.g. breathing, phonatory and articulatory gestures, kinaesthetic control) as well as a multi-level cognitive treatment based on the synchronised activation of various cortical areas (e.g. motor area, perception areas, language areas). Defining speech rhythm thus remains difficult, although it constitutes a fundamental prosodic feature. From this angle, the question of generating "natural" synthetic rhythm remains one of the largely unresolved challenges in speech synthesis. For example, the European COST action 258 — involving about 30 research laboratories in 17 countries — has identified this problem as one of the major issues for the improvement of the naturalness of synthetic voices.

The acknowledged complexity of what encompasses rhythm partly explains that the common approach to describing speech rhythm is based on a few parameters (such as stress, energy, duration), which are the phonologically represented parameters. In this paper, we postulate that the rhythmic poverty of artificial voices is related to the fact that rhythmicity is not sufficiently captured by current phonological models. We argue that our formal "tools" are not powerful enough and that they reduce our capacity to understand phenomena such as rhythmicity. Section 1 in this paper will discuss these issues. As a consequence, rhythmicity, to the degree that it is provided by current phonological models and implemented in European speech synthesisers, tends to be rather poor, even when just a neutral declarative style is generated.

The temporal component in speech rhythm

In our mind, the insufficiencies in the description and synthesis of rhythm are in part related to the larger issue of how speech temporal structure is modelled in current phonological theory. Indeed, prosody modelling has often been reduced to the description of accentual and stress phenomena, and truly temporal issues such as pausing, varying one's speech rate or "time-interpreting" the prosodic structures have not been as extensively examined, and have not yet been formalised in theory. It is claimed that the status of the temporal component in a prosodic model is a key issue for two reasons. First, it enters into the understanding of the relations between the temporal and the melodic components. Second, it enters into the modelling of different styles of speech, which requires prosodic flexibility.

Relations between the temporal and the melodic components

Theoretical modifications are not undertaken lightly. It seems to us that we must start with an acknowledgement of a lack of knowledge: we should not assume that the status of the temporal component in a prosodic system is a solved issue. At the same time, this type of knowledge is important. For example, understanding how the temporal component relates to the melodic component within a prosodic system is clearly of fundamental importance. We need to understand in

detail how to interpret relations between the temporal and the melodic components at the phonological level. Such relations are not yet fully elucidated in current approaches, either from a theoretical or an engineering point of view. This issue is further complicated by the fact that there is no evidence that timing-melody relations are stable and identical across languages (Sluijter, 1995; Zellner, 1996, 1998), or across various speech styles. Similarly, our work on French indicates that the tendency of current prosodic theories to invariably infer timing-melody relations solely from accentual structures leads to an inflexible conception of the temporal component of speech synthesis systems.

Flexible prosodic models

It is likely that for the modelling of different styles of speech (running texts, lists, addresses, etc.) — another upcoming challenge in speech synthesis — current rhythmic models conceived for declarative speech would not be appropriate. Does each speech style require an entirely different rhythmic model, a new structure with new parameters? If so, how would such different rhythmic models be related to each other within the same overall language structure? The difficulty of formalising obvious and coherent links between various rhythmic models for the same language will likely impede the development of a single dynamic rhythmic system for a given language**.** The resulting disjointed view of speech rhythm may thus obscure our understanding of speech rhythmicity, instead of serving to enlighten it.

In summary, we suggest that more *explicitness* in the representation of the features contributing to the perception of speech rhythm would facilitate the scientific study of rhythm. A richer representation would permit a better understanding of speech rhythmicity. Furthermore, if we had such a formalism at our disposal, it would probably become easier to define and understand the exact nature of relations between intonation — *i.e.*, model of melodic contours — and temporal features.

In the following sections, current concepts of speech rhythm will be discussed in more detail. Subsequently, two non-speech human communicative systems will be examined, dance and music notation, since they also deal with rhythm description. These non-verbal systems were chosen because of their long tradition in coding events which contribute to rhythm perception. Moreover, as one system is mainly based on body language perception and the other one is mainly based on auditory perception, it is interesting to look for "invariants" of the two systems when coding determinants to rhythm. Looking at dance and music notation may help us better understand which information is still missing in our formal representations.

1. REPRESENTING RHYTHM IN PHONOLOGY

For speech scientists, generating speech synthesis concretises the link between phonology — the representation of a certain level of linguistic organisation — and the phonic substance — speech as it is captured by acoustical tools. Currently in Europe, prosodic structures are integrated into phonological models, and two principal types of abstract structures have essentially been proposed. *Tonal prominence* is assumed to represent pitch accent, and *metrical prominence* is assumed to represent temporal organisation and rhythm (cf. among others Gussenhoven, 1988; Nespor & Vogel, 1986; Pierrehumbert, 1987; Selkirk, 1984). Rhythm in the metrical approach is expressed in terms of prominence relations between syllables. Selkirk (1984) has proposed a metrical grid to assign positions for syllables, and others like Kiparsky (1979) have proposed a tree structure. Variants of these original models have also been proposed (for example, Hayes, 1995). Beyond their conceptual differences, these models all introduce an arrangement in prosodic constituents and explain the prominence relations at the various hierarchical levels.

1.1 Inflexible Models

These representations are considered here to be insufficient, since they generally assume that the prominent element in the phonetic chain is *the* key element for rhythm. In these formalisations, durational and dynamic features (the temporal patterns formed by changes in durations and tempo) are either absent or underestimated.

This becomes particularly evident when listening to speech synthesis systems implementing such models. For example, the temporal interpretation of the prosodic boundaries usually remains the same, whatever the speech rate. A typical example is the interpretation of a minor prosodic boundary, which is generally realised as a moderate lengthening of the duration of the final syllable. However, Zellner (1998) showed that the "time-interpretation" of the prosodic boundaries is dependent on speech rate, since not all prosodic boundaries are phonetically realised at all speech rates. The analysis of statistically significant groups of durations observed at prosodic boundaries shows two trends related to tempo. At a fast speech rate, the temporal structure tends to be organised around the major temporal breaks (major boundary or pause). At a slow speech rate, the temporal structure tends to be organised around all prosodic boundaries, with no significant distinction between minor and major phrases. A

prosodic model should take into account these different strategies for the realisation of prosodic boundaries.

1.2 Binary Models

Tajima (1998) pointed out that *"metrical theory has reduced time to nothing more than linear precedence of discrete grid columns, making an implicit claim that serial order of relatively strong and weak elements is all that matters in linguistic rhythm"* (p.11). This "prominence approach" shared by many variants of the metrical model leads to a rather rudimentary view of rhythm. It can be postulated that if speech rhythm was really as simple and binary in nature, adults would not face as many difficulties as they do in the acquisition of rhythm of a new language. Also, the lack of clarity on how the strong-weak prominence should be phonetically interpreted leads to an uncertainty in phonetic realisation, even at the prominence level (Coleman, 1992; Local, 1992; Tajima, 1998). Such a "fuzzy feature" would be fairly arduous to interpret in a concrete speech synthesis application.

1.3 Natural richness and variety of prosodic patterns

One may suspect that prominence is not the only determinant of speech rhythmicity. In examining the detailed timing structure of a variety of speech synthesisers of French, we have noticed that their unnatural and mechanical rhythm tended to result from of an excessive number of tonal prominences and a lack of temporal variation and dynamism — *i.e.*, a lack of temporal patterns which are provided by word-groupings and by the various types of "temporal boundaries" (as defined by Zellner 1996, 1998). After hearing one minute of synthetic speech, it is often easy to mentally conjecture what the prosodic pattern of various speech synthesisers will sound like in subsequent utterances, suggesting that commonly employed prosodic schemes are too simplistic and too repetitive. Natural richness and variety of prosodic patterns probably participate actively in speech rhythm, and models need considerable enrichment and differentiation before they can be used to predict a fully natural and fluid prosody for different styles of speech. In that sense, we should probably take into account not only perceived stress, but also the hierarchical temporal components making up an utterance.

This is important not only for the perceived naturalness of synthetic speech, but also to support the perceptual clarity of the transmitted information. Dessons & Meschonnic (1998), define rhythm as an organisation of *meaning* across the alternation of accents, sound effects, and prosodic organisation. In other words, rhythm is seen as a key prosodic tool for signalling an overarching semantic organisation. Some of this can be documented with respect to temporal aspects of the issue. By increasing articulatory speed in certain places of the utterance, and by decreasing it in others, a sense of coherence over an entire semantic unit can be established. But exactly how is this done? And how can we mark up text, either manually or automatically, to enable our synthesisers to implement this aspect of prosodic processing adequately?

The following section indicates how the transition from the phonological formalism to the automatic generation of speech rhythm might be achieved.

2. RHYTHM IN SPEECH SYNTHESIS

Established knowledge suggests that any overall prosodic scheme should begin by incorporating local linguistic effects obtained both from tonal and metrical structures. In fact, it may be profitable to pose the temporal structuring problem in bottom-up hierarchical and multidimensional terms.

For example, at the segmental level, Keller and Zellner (1995, 1996) showed that in a French corpus of 100 sentences, sound segment[1] distribution (*i.e.*, exactly which sound segments make up the utterance) can by itself — with no prosodic information — explain a substantial proportion of the variance of the temporal organisation of speech. This means that for French, intrinsic segment duration, combined with the language- and context-specific sound distribution, is by itself a vital temporal determinant of direct relevance for the perception of rhythm.

It may be interesting to note in this context that the great conductor William Christie claimed intuitively[2] that an essential rhythmic element in French is invested in consonants, contrary to Italian, where vowels carry the principal responsibility for establishing rhythm.

In other words, an initial and crucial step in calculating the overall rhythmic timing scheme is to develop a language-specific and segmentally adequate timing model for segments. Notice that according to the speech rate, segmental durations will change not only in terms of their intrinsic durations but also in terms of their relations within the segmental system since all the segments do not present the same "durational elasticity" (Gay, 1981; Vaxelaire, 1994; Zellner, 1998).

At successively higher levels, we may wish to model syllabic, word-level, small word-group level and major word-group level effects.

[1] Speech units such as the sound segments are easy to manipulate and are frequently-used basic components in current speech synthesis systems. This is the reason why these units are used.

[2] In an interview on French television, spring 1999.

Implementing overall objectives by proceeding from the bottom up, as we have done in several of our own analyses (Keller & Zellner, 1996), has the advantage that each timing effect is modelled at its own relevant level. The massive timing effects that are required for lower-level models are implemented first, leaving only residual and directly relevant effects to be modelled at each successively higher level, until the semantically more relevant levels of the major word group are reached, *i.e.*, the sentence and the paragraph. For example, at the lexical level, certain typical patterns have been established for English and Japanese lexicons: in English, the typical pattern is the stress-foot structure which favours initial stress, and in Japanese, the typical pattern is the bimoraïc foot structure (Tajima, 1998). Such patterns can be implemented as modelling reaches syllable and foot levels.

It thus appears that speech rhythm is a complex multidimensional parameter, which cannot be well represented in terms of a sole strong - weak prominence distinction within a serial sequence. The issue then becomes how to better understand what the cognitive relevant features are, in order to be able to recreate rhythm adequately in synthetic speech.

It may be useful to consider the analysis of rhythm in other domains where this aspect of temporal structure is vital. This may help us identify the formal requirements of the problem. If the first obstacle speech scientists have to deal with is indeed the *formal representation* of rhythm, it may be interesting to look at dance and music notation systems, in an attempt to better understand what the missing information in our models may be.

3. REPRESENTING RHYTHM IN DANCE AND MUSIC

Speaking, dancing and playing music are all time-structured objects, and are thus all subject to the same fundamental interrogations concerning the notation of rhythm. For example, dance can be considered as a frame of actions, a form that progresses through time, from an identifiable beginning to a recognisable end. Within this overall organisation, many smaller movement segments contribute to the global shape of a composition. These smaller form units are known as “phrases” which are themselves composed of “measures” (or “meters”) based on “beats”.

The annotation of dance and music has its roots in antiquity and demonstrates some improvements over current speech transcriptions. This annotation is still the dominant means of codifying temporal structures, despite the more recent availability of video tools and recently-developed computer-based behavioural scoring systems. Video recordings are excellent for giving an impression of the work as a whole, as interpreted by a particular artist. This is invaluable if it is the dancer or the musician who is the object of attention, but it does not capture the essential, more abstract components of the dance or the music. In video recordings, the performer's contribution is superposed on the choreographer’s or the composer’s original intention, while notation is a more scientific record of the intended action, recorded in symbols in order to represent almost every segment.

Even though such notations generally allow many variants — which is the point of departure for artistic expression —, they also allow the retrieval of a considerable portion of rhythmic patterns. In other words, even if such a system cannot be a totally accurate mirror of the intended actions in dance and music, the assumption is that these notations permit a more detailed capture and transmission of rhythmic components.

Mechanically produced music (*e.g.*, a piece played on a player piano or a MIDI synthesiser on the basis of notation alone) is a good illustration that music notation can indeed capture some of these rhythmic features. In these cases, music is generated from paper-rolls made on a punching machine according to music partitions or, in the case of MIDI reproduction, music is generated from an automatic translation of the notes. Although a musically trained ear can easily distinguish a piano player or a mechanical MIDI interpretation from a natural (human) interpretation, some primary vital elements contributing to the recreation of rhythm are undoubtedly present in these representations. The next sections will render more visible these elements by looking at how rhythm is encapsulated in dance and music notation.

3.1. Dance notation

In dance, there are two well-known international notation systems: The Benesh system of Dance Notation[3] and the Labanotation[4]. Both systems are based on the same lexicon that contains around 250 terms. An interesting point is that this common lexicon is hierarchically structured.

A first set of terms designates static positions for each part of the body: legs, arms, body, hands, etc. For example: “cinquième” designates a particular position of the feet.

[3] Benesh system: www.rad.org.uk/index_benesh.htm

[4] Labanotation: www.rz.unifrankfurt.de/~griesbec/LABANE.HTML

Position of the feet: cinquième.
(©1996 AMERICAN BALLET THEATRE)

Attitude.
(©1996 AMERICAN BALLET THEATRE)

"Attitude" designates a more complex position where the dancer stands on one leg with the other raised in back, the knee bent at an angle of 90 degrees and well turned out, so that the knee is higher than the foot. The arm on the side of the raised leg is held over the head in a curved position, while the other arm is extended to the side.

A second set of terms designates patterns of steps that are chained together. These dynamic sequences thus contain an intrinsic timing of gestures, providing a primary rhythmic structure. For example: "ballotté, pas de bourrée, saut de chat, pas chassé". These sequences of actions necessarily take a certain time to execute, and thus intrinsically establish a fundamental timing scheme.

The third set of terms designates spatial information with different references, such as pointing across the stage or to the audience, or references from one to another part of body.

The fourth level occasionally used in this lexicon is the "type" of dance, the choreographic form: a rondo, a suite, a canon, etc.

Since this lexicon is not sufficient to represent all dance patterns, more complex choreographic systems have been created. Among them, a sophisticated one is the Labanotation system which permits a computational representation of dance. Labanotation is a standardised system for transcribing any human motion. Its staff consists of three lines and runs vertically (Figure 1).

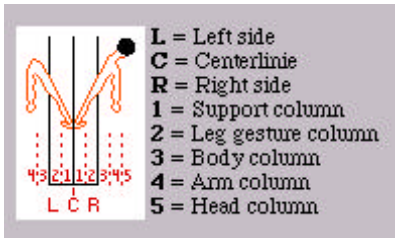
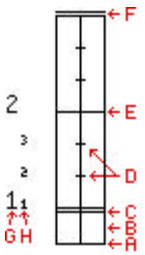


Figure 1. (©1996 Christian Griesbeck, Frankfurt/M)

The score is read from the bottom to the top of the page (instead of left to right like in music notation). This permits noting on the left side of the staff anything that happens on the left side of the body and vice versa for the right side (Figure 2).



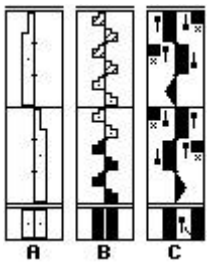
A = Line at the start of the staff
B = Starting position
C = Double line indicates the start of the movement
D = Short line for the beat
E = Bar line
F = Double line indicates the end of the movement
G = Large numbers for the bar
H = Small numbers for the beat (only in the first bar)

Figure 2. (©1996 Christian Griesbeck, Frankfurt/M)

In the different columns of the staff, symbols are written to indicate in which direction the specific part of the body should move. The length of the symbol shows the time the movement takes, from

its very beginning to its end. If nothing special is annotated, all movements should be performed normally. This can be modified by using space measurement signs (1 to 6: from slightly in the first degree to totally in the sixth degree).

To record if the steps are long or small, if the arm or legs are bent or extended, space measurement signs are used. If the movement is accented, 14 accent signs are used (heavy - gentle, strong - shaking, strong resilient - resilient - relaxed, emphasised - unemphasised, uplift - passive, very strong - strong - slight). If a special overall style of movement is recorded, key signatures (e.g. ballet) are used. To write a connection between two actions, Labanotation uses bows (like musical notation). Vertical bows show that actions are executed simultaneously, they show phrasing, include body parts or add specific aspects to the movement. Horizontal bows show a connection with space (Figure 3).



A = Very slow steps (one half step per bar)

B = Very fast steps (one step per quarter note) first in low level - with bend knees (in Plié) like Groucho Marx. Next in middle level - normal steps. In the end in high level - steps on half point.

C = Balancé: The right leg steps to the right side into the Plié. Step on the left leg into the Plié in place in front of the right leg (indicated by the pin) - simultaneously the right leg sur le cou-de-pied back (the small x in front of the leg gesture is a Space measurement sign). Step on the right leg in place into Plié back - simultaneously the left leg sur le cou-de-pied front. In the next bar everything to the other side.

Figure 3. Example with various steps (©1996 Christian Griesbeck, Frankfurt/M)

In conclusion, dance notation is based on a large structured lexicon that contains some intrinsic rhythmic elements (patterns of steps). Some further rhythmic elements are represented in a notation system like Labanotation, such as the length of a movement — equivalent to the length of time, the degree of a movement (the quantity), the accentuation, the style of movement, and possibly the connection with another movement.

3.2. Music Notation

Turning to music, rhythm is clearly identified as one of its prime elements. It affects how long musical notes last (duration), how rapidly one note follows another (tempo), and the pattern of sounds formed by changes in duration and tempo (rhythmic changes). Rhythm in Western cultures is normally formed by changes in duration and tempo (the non-pitch events): it is normally metrical, that is, notes follow one another in a relatively regular pattern at some specified rate.

In music, the standard system used today (5-line staffs, keynotes, bar lines, notes on and between the lines, etc.) was developed in the 1600s from an earlier system called "mensural" notation. This system permits a fairly detailed transcription of musical events.

For example, *pitch* is indicated both by the position of the note and by the clef. *Timing* is given by the length of the note (colour and form of the note), by the time signature and by the tempo. The time signature is composed of bar-lines ("|" ends a rhythmic group), coupled with a figure placed after the clef (e.g., 2 for 2 beats per measure), and below this figure is the basic unit of time in the bar (e.g., 4 for a quarter of a note, a crotchet). Thus, "2/4" placed after the clef means 2 crotchets per measure.

Then comes the tempo which covers all variations of speed (e.g. *lento* to *prestissimo*, number of beats per minute). These movements may be modified with expressive characters (*e.g., scherzo, vivace*), rhythmic alterations (*e.g., animato*) or accentual variations (*e.g., legato, staccato*).

Grave	very slow
Largo	broadly
Larghetto	rather broadly
Lento	slow
Adagio	slow, leisurely
Andante	walking pace
Moderato	moderate time
Allegretto	slightly slower than Allegro
Allegro	lively, fast
Vivace	quick, lively
Presto	very fast
Prestissimo	as fast as possible

The part left out of consideration in the musical notation system for the representation of the temporal structure corresponds to the difference

between a human artistic performance and a strictly mechanical performance.

In summary, music notation is based on a spatial coding — the staff. A spatially sophisticated grammar permits specifying temporal information (length of a note, time-signature, tempo) as well as the dynamics between duration and tempo. These features are particularly relevant for capturing rhythmic patterns in Western music, and from this point of view, an illustration of the success of this notation system is given by mechanical music as well as by the rhythmically adequate preservation of a great proportion of the musical repertoire of the last few centuries.

3.3. Conclusion on these notations

In conclusion, dance notation and music notation have shown that elements which contribute to the perception of rhythm are represented at various levels, reinforcing our initial notion that rhythm is a complex multidimensional parameter.

Please note that there is no intention to claim that such notation systems achieve a perfect representation of rhythm. It is well known that these scoring systems admit a relatively large freedom of personal interpretation. Moreover, intrinsic human timing is not coded in such systems, whether it is produced with vocal or manual gestures. For example, the notation of a *staccato* on a piano contains no implicit timing for finger movement. Certain assumptions about intrinsic execution times of specific actions and action sequences are simply presupposed. But once such suppositions are taken into account, current notation systems do permit a coding of rhythm that permits an approximate reproduction within a framework judged satisfactory in most of the dance and music community.

Furthermore, notice that much rhythmic information is given by temporal elements at various levels such as the "rhythmic unit" (duration of the note or the step), the lexical level (patterns of steps), the measure level (time-signature), the phrase level (tempo), as well as by the dynamics between duration and tempo (temporal patterns). Therefore both types of notation represent much more information than only the prominent or accentual events. Many relevant hints are given by temporal symbols at various levels of the time-structured object.

4. PROPOSAL OF REPRESENTATION OF RHYTHM IN SPEECH

If one accepts the hypothesis that speech rhythm shares with dance rhythm and music rhythm a common set of vital elements, the comparison of notation systems may let us identify some of the missing information in our formal representations of the speech utterance train.

A striking feature in dance and music notations, as shown in the former sections, is the extensive amount of temporal information which is represented for music and dance, but which is typically absent in our phonological models. It is so much more surprising that a great number of timing models have been suggested in speech synthesis for computing the duration of speech units (see for example Barbosa, 1994; Campbell, 1992; de Tournemire, 1998; van Santen, 1993), based on such relatively poor phonological representations. It is true that these models are relatively easy to program and rather promising in terms of the temporal accuracy of low-level speech units (*e.g.*, de Tournemire, 1998), but in terms of their rhythmic structure, they remain poor in the sense that they are easy to predict and barely varied at the empirical level. Listening to synthetic voices generated with these models is still a tiresome and monotonous task. Moreover, much difficulty can be anticipated when this "knowledge" is transferred to a new style of speech or to a new language.

It is thus proposed to enrich our representations of speech. If rhythm perception results from multidimensional "primitives", our assumption is that the richer prosodic formalisms are, the better speech rhythmical determinants will be. In this view, three kinds of temporal information must be retained: the tempo, the dynamic patterns and the durations.

1. How fast syllabic units are produced (tempo): slow, fast, explicit, etc. *Tempo* is given at the utterance level (as long as it doesn't change), and should provide all variations of speed. It seems for example that in many languages, a fast speech rate involves the production of 7 to 8 syllables per second, and a slow speech rate involves the production of 5 to 6 syllables per second (Fujisaki, 1998). If this is confirmed, this could be considered to be a standard to which tempo is pegged.

In our mind, the preliminary establishment of a speech rate in a rhythmic model is important for three reasons.

First, speech rate gives the temporal span by the average number of syllables per second.

Second, in our model, it also involves the *selection* of the adequate intrinsic segmental durational system. Among others, Zellner (1998) showed that this segmental durational system is deeply restructured with the change of speaking rate. As explained in section 2 of this paper, for French, intrinsic segment duration, combined with the language- and context-specific sound distribution, is by itself a vital temporal determinant of direct relevance for the perception of rhythm.

Third, some phonological structurings related to a specific speech rate can then be modelled: for

example in French, schwa treatment or precise syllabification (Zellner, 1998).

2. The dynamics relating various groups of units, *i.e.*, the temporal patterns formed by changes in duration and tempo: word grouping and types of "temporal boundaries" as defined by Zellner (1996, 1998). ***Temporal patterns*** are automatically given at the phrasing level, thanks to a text parser (Zellner, 1998) and are interpreted according to the tempo (global speech rate). For example, for slow speech rate, an initial minor temporal boundary is interpreted at the syllabic level as a minor syllabic shortening, and a final minor temporal boundary is interpreted as a minor syllabic lengthening. This provides the temporal skeleton of the utterance.

3. How long units last: durations for syllabic and segmental speech units. This component is already present in current models. *Durations* are specified according to the preceding steps 1 and 2, at the syllabic and segmental levels.

The incorporation at the phonological level of the three types of temporal information should permit a better modelling and better understanding of speech rhythmicity. Moreover, the interactions between the three types of temporal information are particularly important. For example, an apparently small change in the temporal structure may have a larger effect on the entire temporal structure. This will be shown in the following section.

5. AN EXAMPLE

In this section, the suggested concepts are illustrated with a concrete example taken from French. The sentence is "The village is sometimes overcrowded with tourists".

"Ce village est parfois encombré de touristes."

1. Setting the Tempo: fast (around 7 syllables/s)

Since the tempo chosen is fairly fast, it has previously been shown that when it is relevant, some final schwas may be "reduced" depending on the closeness to a temporal boundary — see next step (Zellner, 1998).

2.a. Automatic Prediction of the Temporal Patterns

Temporal patterns are initially formed according to the temporal boundaries (m: minor boundary, M: major boundary).These boundaries are predicted on the basis of a text parser (*e.g.*, Keller & Zellner, 1998; Zellner, 1996) which is adapted depending of the speech rate (Zellner, 1998).

2.b. Interpretation of the boundaries and prediction of the temporal skeleton

It has been shown that for French, the interpretation of the predicted temporal boundaries depends on the tempo (Zellner, 1998). For example, for fast declarative reading style, three levels are statistically relevant: major final boundary, prepausal final minor boundary and other positions.

"Ce villag(e) est parfois encombré d(e) touristes."
M M

The temporal boundaries are expressed in levels (see below) according to an average syllabic duration (which varies with the tempo). For example, for fast speech rate:

• A final major boundary (level 3) is interpreted as a major lengthening of the standard syllabic duration

• Within the sentence, a prepausal phrase boundary or a major phrase boundary is interpreted at the end of the phrase as a minor lengthening of the standard syllabic duration (level 2).

• Level 0 indicates a shortening of the standard syllabic duration as for the beginning of the sentence.

• All other cases are realised on the basis of the standard syllabic duration (level 1).

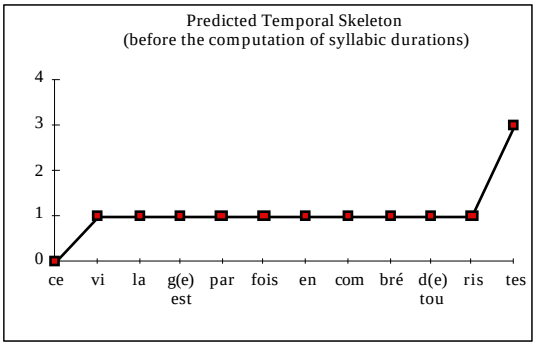


Figure 4. Predicted temporal skeleton for fast speech rate: "Ce village est parfois encombré de touristes."

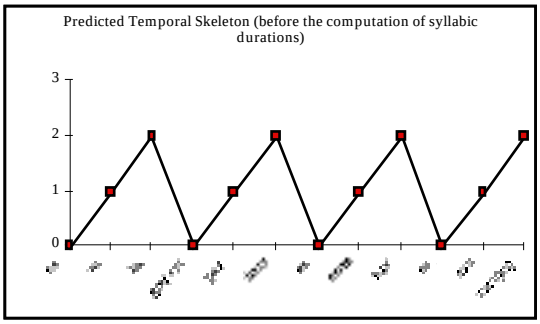


Figure 5. Predicted temporal skeleton for slow speech rate: "Ce village est parfois encombré de touristes."

Figures 4 and 5 show the results of our boundary interpretation according to the fast and to the slow speech rate. Each curve represents the utterance symbolised in levels of syllabic durations. This gives a skeleton of the temporal structure.

3. Computation of the durations

Once the temporal skeleton is defined, the following step consists of the computation of the segmental and syllabic durations of the utterance, thanks to a statistical durational model used in a speech synthesiser. Graphs 6 and 7 represent the obtained temporal curve for the two examples, as calculated by our durational model (Keller & Zellner, 1995, 1996) on the basis of the temporal skeleton. The primitive temporal skeletons are visually clearly related to this higher step.

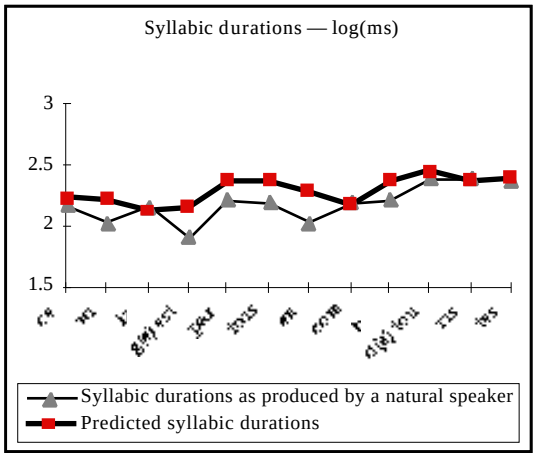


Figure 6. Predicted temporal curve and empirical temporal curve for fast speech rate: "Ce village est parfois encombré de touristes."

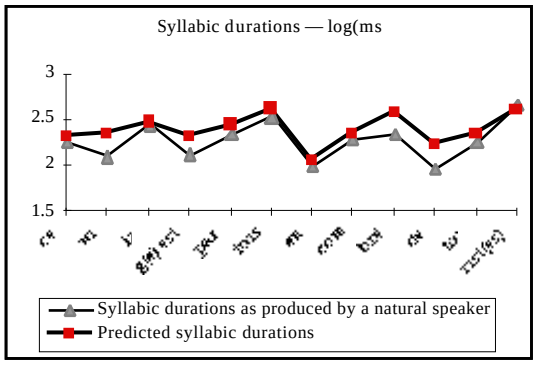


Figure 7. Predicted temporal curve and empirical temporal curve for slow speech rate: "Ce village est parfois encombré de touristes."

Figures 6 and 7 show the vicinity of the predicted curves to the natural ones. Notice that the sample utterance was randomly chosen from 50 sentences.

This example shows to what extent combined changes in tempo, temporal boundaries, and durations impede the whole temporal structure of an utterance which may indeed affect the rhythmic structure. It is thus crucial to incorporate this temporal information in our notations to improve the comprehension of speech rhythm.

The final step in the attempt to understand speech rhythm would involve the comparison of those temporal curves with traditional intonational contours. Since the latter are focused on prominences, this comparison would illuminate the relationship between prominence structures and rhythmic structures.

CONCLUSION

Rhythmic poorness of artificial voices is related to the fact that determinants of rhythmicity are not sufficiently captured with our current phonological models. It was shown that the representation of rhythm is in itself a major issue.

The examination of dance notation and music notation suggests that rhythm coding requires an enriched temporal representation. The present approach offers a general, coherent coordinated notational system. It provides a representation of the temporal variations of speech at the segmental level, at the syllabic level and at the phrasing level (with the temporal skeleton). In providing tools for the representation of essential information that has till now remained under-represented, a more systematic approach towards understanding speech rhythmicity may well be promoted. In that sense, such a system offers some hope for improving the quality of synthetic speech. If speech synthesis sounds more natural, then we can hope that it will also become more pleasant to listen to.

ACKNOWLEDGEMENTS

Our grateful thanks to Jacques Terken for his stimulating and extended review. Cordial thanks go also to our colleagues Alex Monaghan and Marc Huckvale for their helpful suggestions on an initial version of this paper.

REFERENCES

Barbosa, P. A. (1994). *Caractérisation et génération automatique de la structuration rythmique du français.* Thèse de Doctorat. U.R.A. CNRS n° 368 - INPG/ENSERG, Université Stendhal, Grenoble.

Campbell, W.N. (1992). Syllable-based segmental duration. In G. Bailly, & al. (Eds.), *Talking Machines. Theories, Models, and Designs* (pp. 211-224). Elsevier Science Publishers.

Coleman, J. (1992) "Synthesis by rule" without segments or rewrite-rules. In G. Bailly & al. (Eds.), *Talking Machines. Theories, Models, and Designs* (pp. 43-60). Elsevier Science Publishers.

Dessons, G., Meschonnic, H. (1998). Traité du rythme. Dunod (Paris).

de Tournemire, S. (1998). Automatic detection of intonation using an identified prosodic alphabet. *Proceedings of ICSLP, 5th International*

Conference on Spoken Language Processing. Paper 1035, December 1998, Sydney (Australia).

Fujisaki, H. (1998). On the effects of speech rate upon paremeters of the command-response model for the fundamental frequency contours of speech. *Proceedings of ICSLP, 5th International Conference on Spoken Language Processing*. Paper 935, December 1998, Sydney (Australia).

Gay, T. (1981). Mechanisms in the control of speech rate. *Phonetica 38.* 148-158.

Gussenhoven, C. (1988). Adequacy in intonation analysis: the case of Dutch. in Smith & Van der Hust (Eds), *Autosegmental Studies on Pitch Accent* (pp. 95 - 121). Foris, Dordrecht.

Hayes, B. (1995) *Metrical stress theory: Principles and case studies*. (Univ Chicago, Chicago).

Keller, E. (1994). The fundamentals of phonetic science. in E. Keller (Ed.), *Fundamentals of Speech Synthesis and Speech Recognition* (pp. 5-21). Chichester: John Wiley.

Keller, E., & Zellner, B. (1995). A statistical timing model for French. *XIIIème Congrès International des Sciences Phonétiques, 3* (pp. 302-305). Stockholm, (Sweden).

Keller, E., & Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics*, 17, University of York. 53-75.

Kiparsky, P. (1979). Metrical structure assignement is cyclic. *Linguistic Inquiry, 10*. 421 - 441.

Local, J.K (1992). Modelling assimilation in a non-segmental, rule-free phonology. In Docherty, G.J. and Ladd, D.R. (Eds), *Papers in Laboratory Phonology*, II (pp.190 - 223). Cambridge: CUP.

Nespor, M., & Vogel, I. (1986). *Prosodic Phonology.* Dordrecht: Foris.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. Cambridge, MA. MIT Press.

Port, R., Cummins, F., & Gasser, M. (1996). A dynamic approach to rhythm in language: Toward a temporal phonology. In B. Luka and B. Need (eds), *CLS-31: Proceedings of the Chicago Linguistics Society*. (Chicago Linguistic Society) pp. 375-397.

Selkirk, E.O. (1984). *Phonology and Syntax: The Relation between Sound and Structure.* MIT Press, Cambridge, MA.

Sluijter, A. M. C. & van Heuven, V. J. (1995). Effects of focus distribution, pitch accent and lexical stress on the temporal organisation of syllables in Dutch. *Phonetica, 52.* 71-89

Tajima, K. (1998). *Speech Rhythm in English and Japanese: Experiments in Speech Cycling.* Ph. Dissertation. Indiana University.

van Santen, J.P.H (1993). Timing in text-to-speech systems. *Proceedings of the 3rd European conference on speech communication and technology* (pp. 1397-1404). Berlin.

Vaxelaire, B. (1994). Variation de geste et débit. Contribution à une base de données sur la production de la parole, mesures cinéradiographiques, groupes consonantiques en français. *Travaux de l'Institut de Phonétique de Strasbourg, 24.* 109-146.

Zellner, B. (1996). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée: La communication parlée.* 1. (pp.7-23). Paris.

Zellner, B. (1996). Relations between the temporal and the prosodic structures of French, a pilot study. *Proceedings of Annual Meeting of the Acoustical Society of America,* Honolulu, HI, USA.

Zellner, B. (1998). *Caractérisation et prédiction du débit de parole en français. Une étude de cas.* Thèse de Doctorat. Faculté des Lettres, Université de Lausanne.