MOTIVATIONS FOR THE PROSODIC PREDICTIVE CHAIN

Eric Keller and Brigitte Zellner

Laboratoire d'analyse informatique de la parole (LAIP), IMM, Lettres
University of Lausanne, 1015 Lausanne, Switzerland
Eric.Keller@imm.unil.ch and Brigitte.Zellner@imm.unil.ch

ABSTRACT

Prosodic modelling is presented as a two-part process. In the first part, predictor variables are identified on the basis of psycholinguistic factors that determine word grouping. In the second part, these and intrinsic-contextual linguistic factors generate predictions for durational and melodic assignment. It is argued that data-driven approaches can be rendered more efficient by introducing and respecting psycholinguistic principles of human prosodic processing. This study summarises thinking developed conjointly in our laboratory since 1993, as well as portions of the second author's thesis (Zellner, 1998).

1. INTRODUCTION

Constructors of speech synthesis systems are confronted by a wide and sometimes bewildering set of theories concerning prosodic processing. On the one hand, there are fairly abstract linguistic accounts, generally conceived for small portions of text in a specific language, which often prove difficult to implement in another language and with general applicability to any type of text. On the other hand, there are data-driven approaches which, laboriously and using fairly gigantic data sets, seek to infer prosodic realities directly from target speech, without following a clear set of theoretical guidelines.

In our experience, both of these common approaches have some major failings. For example, we were unable to develop a satisfactory extension of the most common linguistic accounts of word grouping to French (see section 2). Also we would expect that a strict data-driven account would require the manipulation of too many unknown variables to be an efficient predictor of word group structure.

In constructing the prosodic component of LAPIS, the Lausanne speech synthesiser, we have thus been guided by a third approach. We have systematically applied the principle that with respect to the initial aspects of prosodic processing, a synthesiser should attempt to capture the functioning of psycholinguistic processes underlying human speech behaviour. This orientation has permitted us to create a system that deals successfully with just about any text in French. Also, it let us identify a set of apparent psycholinguistic constants that simplify and focus data-driven approaches in further prosodic discovery and testing.

2. LINGUISTIC APPROACHES

An initial task of a prosodic model is to identify groups of words or "prosodic phrases". Concretely, this means the identification of words in a speech chain that are considered to be more strongly linked to each other than others.

In linguistic approaches to this issue, intra-word cohesion is generally assumed to be constrained by prosodic structures, which are in turn derived from the interaction between syntactic and phonological structures — at least in languages where syntactic functions are communicated by means of word order.

When this postulate is taken beyond the diversity of theoretical approaches, it implies for French that the presence of word group boundaries should be marked by phrase stresses or phrase accents. As a result, algorithms that calculate the temporal structure in this manner, take the accentual unit as the base of their temporal calculations (Barbosa, 1994; Beaugendre, 1994; Pasdeloup, 1992). In this line of approach, the duration of syllables preceding the accented syllable is a function of the distance from the base unit. The closer the syllable is to the base unit, the greater should be its relative duration.

However, our observations on French are in contradiction with the presumed priority of phonosyntactic and accentual structures. In addition to the largely unresolved problem of assigning unquestionable phrasal stress in French (see Zellner, 1998), there exists to our knowledge no valid theoretical argument that motivates the derivation of temporal structures from phonosyntactic or accentual structures in languages such as French (for a detailed argumentation, see Zellner, 1998). Moreover, no agreement emerges from the acoustic or the perceptual literature concerning the empirical definition of accent in French (Astesano & al., 1995; Behne, 1989; Pasdeloup, 1992; Zellner, 1998). For these reasons, we found it appropriate to reconsider the basic foundations of the prediction of the temporal structure, and to consider some psycholinguistic bases of word grouping and temporo-melodic assignment instead.

3. THEORY OF PROSODIC CODING IN SPEECH SYNTHESIS

A prosodic component as used in a typical speech synthesis system can be seen to operate in two major phases (Figure 1). During a first phase of the predictive process, and as indicated above, word groups must be constituted to form a structured framework of so-called "prosodic phrases". Once constituted, the framework permits to assign places definable in terms of their proximity to various phrase, word and syllable boundaries to constituent words, syllables and segments. Also during this first operational phase, a set of intrinsic and contextual values must be identified concerning each constituent word, syllable and segment. By "intrinsic characteristics" is meant, in the case of segments, aspects such as their phonetic value (e.g., an [a]

rather than an [i]), or their typical duration, and in the case of words, their grammatical or lexical function. Syllables can also have intrinsic values, such a specific structure. By "contextual information" is meant, particular conditions that are set up by the proximity of sets of segments, syllables and words.
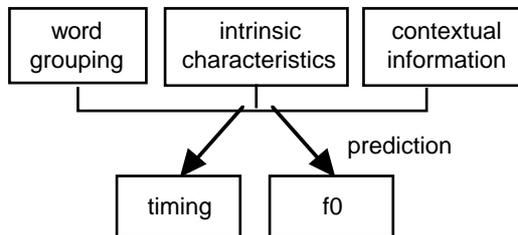

Figure 1. The fundamental structure of prosodic prediction in our system.

In the second stage, relevant positional, intrinsic and contextual information is conjointly used to predict timing and f0. In this phase, it is commonly found that differences in intrinsic or contextual value, or in word, syllable or segment position, translate directly into statistically significant differences in timing and f0.

Operationally, the identification of word groups (which in most traditional systems is handled by some form of a phonosyntactic system) is in our system derived from psycholinguistic principles. In this paper, we wish to argue that a psycholinguistic framework has proven both simple and productive for this first algorithmic phase of prosodic prediction, at least with respect to a neutral reading of declarative French text.

The identification of intrinsic and contextual values, on the other hand, proceeds as in most other systems, by simple inspection of the phonetic chain, or by dictionary lookup.

For the prediction process, either statistical general linear modelling or neural network training can be used. As many other authors, we employ the more explicit method of general linear modelling, since the increased benefits of neural network modelling have proved minimal (e.g., Keller & Zellner, 1995, 1996; Zellner, 1998). In the case of durational prediction, we proceed by direct calculation of segmental duration while taking into account current positional, intrinsic and contextual factors affecting the enclosing segment, syllable and word. In the case of f0 prediction, we initially employ the durational structure to situate an initial, neutral, but speaker-adapted Fujisaki curve, and then we use the segment-specific f0 prediction to adjust exact f0 values via modifications of the "accent commands" (we have implemented "accent commands" for every syllable).

It is to be noted that we use two separate predictive paths for the determination of f0 and duration, while other authors quite commonly predict duration from f0. This is because we did not find particularly strong correlations between the two prosodic measures in our data. For an explicit example, see Figure 2. Similarly low correlations can also be found in much larger French-language data sets.

The crucial innovation in our system is therefore the replacement of a phonosyntactic analysis by a psycholinguistically-derived word grouping algorithm. The major benefit of our procedure is to permit us to operate with only very "light" proximal grammatical processing. This is to disambiguate a relatively small number of homographic conditions, such as "président", which are pronounced differently when the meaning is "le président" or "ils président". This is what we turn to next.
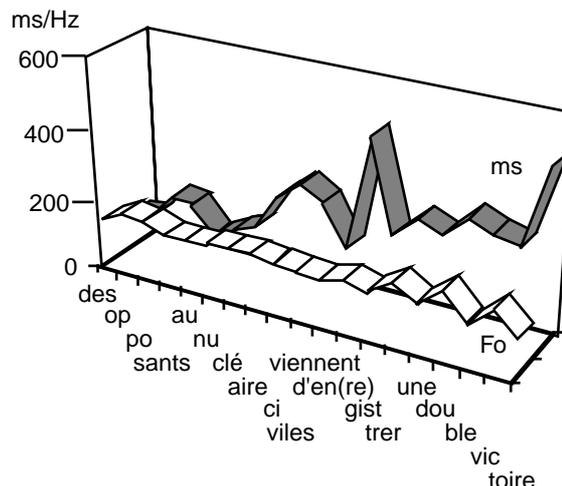

Figure 2. Evolution of syllable duration and syllable-nucleus f0 for a sentence taken at random from a French newscast: «Des opposants au nucléaire viennent d'enregistrer une double victoire.» ["Opponents to nuclear power have just registered a double victory"]. Anchor man, Antenne 2 (Pearson *r*=-0.07).

4. PROSODY AS HUMAN SPEECH PROCESSING

A psycholinguistic account of the human implementation of prosodic prediction must flow from and integrate with a wider model that formalises the real-time process of speech generation. Levelt's (1989) model is probably the most explicit psycholinguistic account of speech production (Zellner, 1998). It integrates the results of a vast number of psycholinguistic experiments and investigations into a single, coherent model. Levelt suggests that three major processing components are active during the generation of speech (pp. 9):

a. A "conceptualiser" that generates the pre-verbal components (if you like, the "conceptual chunks") of a complete message,

b. A "formulator" that contains three phases: grammatical encoding, phonological encoding, and phonetic encoding,

c. An "articulator" that translates the phonetic plan into commands to the phonatory musculature.

Within this framework, prosodic processing is primarily part of the "formulating" stage, with the important specification that "formulation" is not seen to proceed in complete sentences or utterances, but in chunks of smaller groups of words (although concurrent larger-scale planning is likely to take place). According to the psycholinguistic rationale presented by Levelt,

the order of grammatical encoding, as well as the size of the word group, depends in part on the order on which the "lemmas" (parts of the lexicon) become available to the speaker in the first place, an order which itself depends on the order in which the various fragments of the utterance are required by the speaker ("Fluent speech requires incremental production", Levelt, 1989, p. 245).

In this view, the coding of the temporal structure can be seen to typically proceed in terms of a parallel cascading processing of "sets of elementary words", i.e., from one word group to the next. Material is assembled for outputting by the "formulator" in three phases, proceeding from metric planning at the level of each phonological word, via the creation and metric structuring of phonological groups of words, to the metric planning of the entire intonational group.

Levelt also argues that phrase grouping depends on speech rate and on the tendency to produce phrases of relatively well-balanced length, as established by Grosjean and his colleague (Grosjean & Dommergues 1983). This latter tendency might find its origin in motor aspects of human speech. It may be associated either with a buffering limit for speech motor commands and/or with the quantity of air that can comfortably be inhaled for the production of any one speech group.

It is interesting to note that neurophysiological investigations lend further support to this general account of real-time speech processing. According to current neuropsychological hypotheses, the brain processes language in terms of an interaction of three major neuronally-based structures (Damasio et al., 1997). A first structure constitutes representational forms, a second structure processes words and word groups linguistically (this involves the traditional linguistic cortex, i.e., Wernicke's and Broca's areas), and a third structure mediates between the two levels of resprensentation by a set of coordinating and linking techniques, such as word selection and associated processes. Within this general process, it is interesting to consider some of the directly relevant details of the real-time human assembly of word groups.

5. THE CREATION OF WORD GROUPS

It is important to reiterate that so far, we have only implemented a neutral, declarative style of French. Happily, this language and this language style show a great number of regularities. As a result, our prosodic model required only *some* of the many psycholinguistic concepts that were moulded into a complete production model by Levelt (1989). The two concepts we needed above all were "chunking" and "equilibrium-seeking". We are fully aware that other languages and other language styles must necessarily reach beyond these two relatively simple concepts in order to handle prosodic processing successfully. But for us, this would be no reason to look beyond psycholinguistic models of speech production. We estimate, for example, that Levelt's modelling (particularly his "incremental lexical stimulus" implementation of grammatical formulation) will provide a rich source for further, and more extensive, modelling of more complex styles of speech and/or prosodic modelling in other languages.

In our model, the word grouping process proceeds by assembling words into minor groups, minor groups into major groups, and major groups into paragraphs. The assembling of

words into minor groups is based on extensive psycholinguistic experimentation. It has been demonstrated by means of a variety of tasks: by repetition tasks, by subjective segmentation tasks, by empirical measures of syllabic duration, and by the analysis of pause placement (Grosjean & Dommergues, 1983). Across tasks, subjects have shown a tendency to constitute, or leave intact, the same types of groups in the same textual material. Furthermore, several authors, including Grosjean & Deschamps (1975), have shown that pause occurrence and pause duration are strongly correlated with degree of interlexical cohesion. In French, these word groups corresponded, with a few specific exceptions, to sets of one or several grammatical (function) words, followed by one or several lexical (content) words.

That, of course, can never be the whole story. To make this schema work in a real-world speech synthesis for French, a number of important exceptions must be taken into consideration. First, the status of "grammaticality" is subject to contextual influence. A word like "fact", although generally used as a lexical word ("this is a *'fact*"), can be prosodically assimilated to grammatical word usage in certain contexts ("if in *fact* you think of the **'following**"). Many other words show floating grammatical/lexical status as well, depending on context. A good number of these are frequent enough to have to be handled by special sets of rules. Furthermore, assignment to grammatical status is linked to frequency of usage and to the separability of a multi-word expression. Groups of words, such as fixed expressions (so-called "idioms"), negational expressions ("n'est-ce pas"), complex verbal and adverbial expressions ("*il me l'a bien* 'donné" [he definitely 'gave it to me]), must be processed as "grammatical units" to permit realistic prosodic predictions in the melodic and temporal domains (for details, see Zellner, 1996). Once these and several other types of verbal material are correctly classified, it becomes relatively easy to devise automatic algorithms that identify small groups or words (the "minor" groups) with great reliability (i.e., with excellent predictive capacity for durational modification, as documented in Zellner, 1998; see also Zellner, this volume).

But what are the "major groups"? And if so, how many layers of higher group cohesion must we postulate? In reading, we have found evidence for exactly two more levels of word grouping, the level we call "major groups" and the level of the "paragraph". "Major group" boundaries seem to have two principal origins, the presence (or supposed presence) of punctuation marks, and the "equilibrium principle".

The issue of punctuation raises the larger issue of speech style. Most current speech synthesis attempts to imitate good human performance on a rather specific speech task, which is reading. In humans, this corresponds to a particular style of speech, since a good human reader can relatively easily predict prosodic parameters from the visual appearance of the text to be read, which is not the case when producing spontaneous speech. Also, to a good reader, the presence of punctuation in the text simplifies the problem of identifying major word groups within an overall sentence. In our model, we thus assign major word group breaks whenever we encounter a comma or an appositional hyphen. And further major breaks are inserted on the basis of the "equilibrium hypothesis" (Zellner, 1996; 1997).

"Equilibrium" is a psycholinguistic effect that shows considerable influence in our data set. Quite a number of authors have demonstrated that speakers tend to form minor words groups that are relatively well-balanced, no matter whether word

group length is calculated in terms of number of words or number of syllables (Gee & Grosjean, 1983). A similar tendency is also observed for major word groups, where a tendency to "elevate" a minor group boundary to a major group boundary is observed to occur close to the middle of a lengthy sentence, in our data set after some 12-15 syllables (Delais, 1995; Monnin & Grosjean 1993; Pasdeloup, 1992; Zellner; 1996, Zellner, 1997). Importantly, we have found no association between syntactic factors (e.g., the major syntactic break between an NP and a VP) and major pauses of this type—quite the contrary, the syntactic status of preceding and/or following words seems to bear no relation to the presence of an empirically determined group break (Zellner, 1997).

Finally, we implement a differentiation between sentence and paragraph breaks. The synthesiser lowers f0 significantly more, and pauses are made considerably longer, after the end of a paragraph than after a simple sentence break.

This, then, is the essence of our word grouping model for a neutral reading style for declarative French text: sets of linearly structured minor word groups that are regularly interrupted by major word group boundaries, marked by punctuation or calculated in terms of the "equilibrium" hypothesis. And at the end of a paragraph, a particularly salient break. We assume this to be the basic structural mechanism, to be enriched by further mechanisms to handle semantic focus and other phenomena. This information, together with intrinsic and contextual information, suffices to provide predictor variables in phase 1 of the prosodic prediction process of f0 and duration in a neutral declarative reading of French.

What penalty, if any, is associated with the implementation of this word grouping algorithm? Does this relatively simple structural algorithm introduce particularly frequent mispronunciations and/or prosodic imprecisions? A recent evaluation of the text-to-phoneme translation capacity of seven French-language speech synthesisers in which LAPIS participated (Yvon *et al.*, 1998) showed no evident mispronunciation penalty when the overall percentage of phonetic errors is evaluated. All synthesisers made their share of errors, and ours was no exception. Although our synthesiser appears to have by far the "lightest" grammatical analysis of all evaluated systems, our system did not show any excessive weakness in this respect. In terms of the prosody proper, it can be argued that certain semantic nuances are still neglected in the current version of our algorithm. That may indeed be true. On the other hand, our internal testing and countless public demonstrations suggest that the psycholinguistically-based word grouping algorithm also lends our system an particularly natural speech rhythm, which largely compensates for such presumed deficiencies. We have included a few synthesis examples, to permit listeners to make up their own minds. We furthermore expect that future versions of our algorithm will handle such cases with increasing proficiency.

6. SUMMARY

Prosodic modelling for neutral, declarative French speech synthesis was implemented in LAPIS, our speech synthesis device, as a two-stage model. In the first stage, word grouping, intrinsic and contextual information is assembled into a set of predictor variables. In the second stage, separate predictions are calculated for f0 and duration. Intrinsic and contextual information is available through simple inspection and lexical lookup. However, for word grouping, a special mechanism is required. In contrast to many other systems, we have developed a system based on the two major psycholinguistic concepts of "chunking" and "equilibrium". This provided excellent predictions of empirically-determined minor and major word group boundaries, in contrast to syntactically-derived word grouping which often provided word group breaks that were incompatible with empirical evidence obtained from measures of syllable duration and pause placement.

SOUND FILES

The following texts were generated with LAIPTTS and the algorithm described here:

R76_01: Je me demande si vous savez vraiment parler, ou si vous avez encore des choses à apprendre en prosodie, en syntaxe ou en sémantique.

R76_02: Bonjour! Ceci est un exemple de synthèse de la parole, avec une intégration de tous les paramètres prosodiques.

ACKNOWLEDGEMENTS

REFERENCES

Astesano, C., Di Cristo, A. & Hirst, D.J. (1995). Discourse based empirical evidence for a multi-class accent system in French. *XIIIème Congrès International des Sciences Phonétiques, 4* (pp. 630-633). Stockholm.

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.

Beaugendre, F. (1994). Une étude perceptive de l'intonation du français. Thèse de Doctorat en Sciences de l'Université Paris XI. LIMSI n°94 - 25.

Behne, D. M. (1989). *Acoustic effects of focus and sentence position on stress in English and French.* Ph. D. Thesis, University of Wisconsin at Madison.

Damasio, A. & Damasio, H. (1997). Le cerveau et le langage. *Pour la Science: Les langues du monde.* 8 - 15.

Delais-Roussarie, E. (1995). *Pour une approche parallèle de la structure prosodique*: étude de l'organisation prosodique et rythmique de la phrase française. Thèse de Doctorat, Université de Toulouse-Le Mirail.

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

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

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

Keller, E. (1997). Les théories de la parole dans l'éprouvette de la synthèse. In E. Keller, & B. Zellner (éds.), *Les défis actuels en synthèse de la parole*e, *Études des Lettres, 3*. (pp. 9-27). Université de Lausanne.

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

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

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

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

Pasdeloup, V. (1992). Durée intersyllabique dans le groupe accentuel en français. *Actes des 19émes Journées d'Etudes sur la Parole*. (pp. 531-536). Bruxelles.

Yvon, F., Boula de Mareuil, P., d'Alessandro,C., Auberge, V., Bagein, M., Bailly, G., Bechet, F., Foukia, S., Goldman, J.P., Keller, E., Pagel, V., Sannier, F., Veronis, J., O'Shaughnessy, D., Zellner, B. (1998). Objective evaluation of grapheme to phoneme conversion for Text-To-Speech synthesis in French. *Computer Speech and Language*. Special Issue on Evaluation, Vol 12, No. 3.

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. (1997). Improving speech fluency in French through psycholinguistic principles. *14th CALICO Annual Symposium,* (CD) ISBN 1-890127-01-9. New-York.

Zellner, B. (1997). Fluidité en synthèse de la parole. In E. Keller, & B. Zellner (Eds.), *Les défis actuels en synthèse de la parole*e, *Études des Lettres, 3*. (pp. 47-78). Université de Lausanne.

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.

Zellner, B. (1998). Temporal structures for fast and slow speech rate. This volume.