Keller, E., Zellner, B., & Werner, S. (1997). Improvements in prosodic processing for speech synthesis. Proceedings of *Speech Technology in the Public Telephone Network: Where are we Today?* Rhodes, Greece. September 1997.

IMPROVEMENTS IN PROSODIC PROCESSING FOR SPEECH SYNTHESIS

Eric Keller[1], Brigitte Zellner[1] and Stefan Werner[2]

[1] LAIP, Lettres, University of Lausanne, 1015 Lausanne, Switzerland (eric.keller@imm.unil.ch)

[2] Linguistics and Phonetics, Joensuu University, FIN-80101 Joensuu, Finland (werner@phon.joensuu.fi)

ABSTRACT

For the synthesis of French, a separate modelling of fundamental frequency and timing seems more appropriate than derived modelling. Satisfactory declarative prosody in synthesis can be obtained with a version of Fujisaki F0 modelling that takes into account syllable amplitude differences as well as microprosody. Statistical methods based on both segmental and lexical input provide a satisfactory initial rendering of timing.

1. INTRODUCTION

Much current work on synthetic speech is concerned with improvements whose aim is a better comprehension of synthesised messages. Generally speaking, three types of improvement are possible. First, a system must avoid gross reading errors. Second, it must supply human-like intonation and timing ("prosody"). Third, a system must provide a close approximation of human voice quality. Over the past five years, we have examined prosodic modelling from both a theoretical and an experimental perspective. This report presents some factors we have found to be of importance.

2. SEPARATION OF F0 AND TIMING MODELS

In our survey of the literature of speech synthesis prosody, it was found that most authors assign a secondary position to timing, concentrating instead on intonation (Zellner, 1996). Particularly for English, the dominant procedure is to identify stressed syllables, and to raise their fundamental frequency (F0) and to lengthen their durations. Conversely, F0 is lowered and durations are reduced for unstressed syllables. In the sentence "This was really *well* done", the word "well" is perceived with high stress, and is assigned a high F0 as well as lengthened syllable duration, while "done" is considered to have low stress, and is assigned a low F0 and shortened syllable duration. This follows the observation that in lexical stress languages like English, a correlation holds between perceived stress, F0, and syllable duration.

However, the correlation is neither perfect nor universal. Languages like French are much less characterised by stress than English or German, since it is not distinctive (Dauer, 1983). There is no clear agreement on where stress (or "accent") is perceived in French (Pasdeloup, 1992; Astesano, C., Di Cristo. & Hirst., 1995), and a correlation with timing is neither evident nor easy to demonstrate (see F0-duration data taken from two random French sentences, Figure 1; for details of argument, see Zellner, 1996). As a consequence, separate modelling of F0 and duration seems more appropriate. After considerable examination of alternatives, we opted for Fujisaki modelling in the case of F0, and for a statistical prediction model in the case of duration.

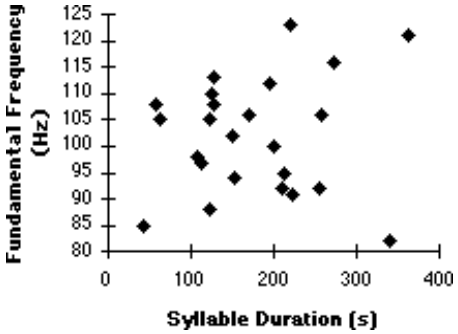

Figure 1. Relationship between Syllable Duration and Mid-Syllable Frequency in two French Declarative Sentences

3. FUJISAKI MODELLING

Quite a few predictive techniques have been developed for the purpose of generating pitch (or F0) contours. They vary in complexity and suitability for different languages. Just four major techniques are distinguished here, (1) phonological prediction schemes, (2) perceptually-derived F0 patterns, (3) neural-network derived F0 patterns, and (4) physiologically-inspired mathematical modelling of F0 contours. Further techniques as well as hybrid solutions are promoted.

For a language like English, a *phonological approach* such as Pierrehumbert (1980, 1981) initially appears promising. Intonation is represented as s sequence of *low* and *high* tones that can play specific roles such as "pitch accents", "phrase accents" and "junction accents" (at phrase boundaries). These units are translated into F0 values on the basis of context-sensitive rules that are passed over the underlying phonological structure from left to right. In addition, there are "downstep rules" that combine with a downwards slant of the F0 base line to produce the typical frequency drop-off over a prosodic phrase.

This model needs relatively few symbols to represent intonation. However, this is also one of its limitations, since a speech signal generated on the basis of such a system shows auditorily significant differences from normal speech signals. First, measured F0 values are not easily assigned binary high-low status, since they often involve intermediate values. Furthermore, each syllable has a characteristic F0 low-high-low contour, called "microprosody" (Figure 2), which is left aside in such a predictive scheme. Finally, a scheme adequate for a

language like English may not always work in other languages. French is probably better described as a boundary-based language than as an accent-based language (Vaissière, 1991), and an accent-based predictive scheme for intonation contours is thus suspect in French.

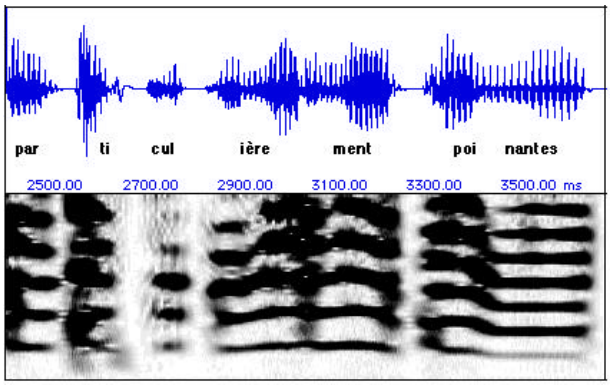

Figure 2: Signal, F0 and Harmonics 1-5 of a French pronunciation of the last part of the sentence "Certaines théories anti-intellectuelles submarginales *sont particulièrement poignantes*". "Microprosody" is visible as a low-high-low tendency in most syllables.

The next two pitch generation schemes are representative of empirical approaches. Here, the effort focuses on the identification of typical and/or minimal intonational contours that characterise speech sequences. Once identified and associated with syntactic and phonological sequences, these patterns can be stored and made available for sequential implementation in a pitch contour.

The intonation model developed at the Dutch Institute for Perception Research (IPO) (see review in 't Hart et al., 1990, and Beaugendre, 1993, for French) aims to incorporate only the *perceptually relevant elements* of an intonational contour. A model for a given language is based on an extensive series of perceptual evaluations of successively more stylised F0 patterns. Once identified, distinctive patterns are stored in terms of three parameters, which are: start and end level, and steepness. For reinsertion into an F0 contour, patterns are aligned on a sentence by means of guidelines that slope downwards from left to right.

There are two major difficulties with this approach. For one, this is clearly a time-consuming and labour-expensive approach to the problem of identifying the F0 contours of a language. This type of effort is thus not easily undertaken to create the intonational data base for a new language. Secondly, it is possible that the technique suppresses information that may not be judged to be perceptually distinctive, but that in fact helps to improve the quality, and thus the degree of naturalness of a synthesised stretch of speech. Microprosodic modulations, for example, are not captured by this approach; nevertheless, their modelling appears to contribute to improved naturalness.

A less labour-intensive means of obtaining F0 patterns was described by Traber for German (1991, 1992). Here, relationships are identified by means of a *neural network* between, on the one hand, sentences described in terms of syllable-specific features and their F0 contours. The features that were retained for the prediction of the F0 contour were syllable accent, type of phrase containing the syllable, presence of a lexical boundary, presence of a sentence-final condition, and syllable position with respect to the main accent. Further features captured the length of the syllable's vowel, high or low intrinsic F0 values, and the voiced/unvoiced quality of the consonantal context. Finally, the model was rendered sensitive to preceding F0 conditions.

Both statistical and perceptual results were considered satisfactory, though Traber observes that rare patterns are sometimes ill-identified by the neural network technique. The approach is certainly intriguing. All of the required predictive features can be obtained mechanically from input text. The technique is straightforward and may well be applicable to new languages without too much difficulty, as long as adjustments are made to compensate for differences in the use of stress.

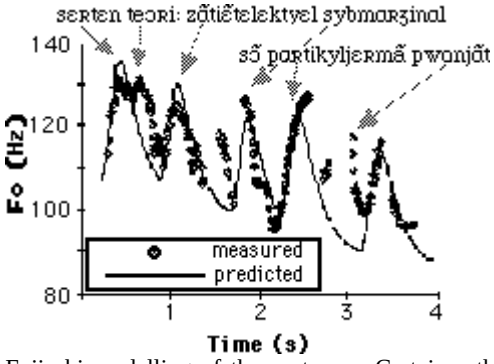

Figure 3. Fujisaki modelling of the sentence «Certaines théories anti-intellectuelles submarginales sont particulièrement poignantes.»

The final approach considered here is the *physiologically-based model* developed by Fujisaki and his colleagues (see e.g., Fujisaki & Hirose, 1982). This model captures two main contours observed to interact in a large number of languages, a global intonational low-high-low contour extending over the entire prosodic phrase and a set of syllable-wide low-high-low contours (Figure 3). Both types of contour are considered to have a physiological basis in typical expiratory speech patterns. The syllable-wide contours are most prominent during stressed syllables in languages like English that show extensive stress modulations; they are in evidence to various degrees during every syllable in those languages like French that show less stress modulation. Syllable-level contours are superimposed on the phrase-level contours to produce a complete approximation to the original contour. To calculate an intonational contour in this fashion, only a few speaker-specific and language-specific parameters are required. These can be obtained by effecting a set of successive approximations between the simulated and the natural contours (Werner, 1995).

The model makes reasonably good approximations for the declarative sentences of a number of languages. In our algorithm for French, we diverged from the Japanese implementation in that we generate syllable-level contours for every syllable of a phrase, rather than just for stressed syllables. We thus generate an approximation to the microprosodic F0 structure. These contours can be generated in real-time on desktop computers.

The only serious downside to Fujisaki modelling we have found is the inevitability of the down trend which does not allow for a natural modelling of all types of interrogative contours. There may also exist some further limits with respect to F0 patterns found in interactive speech. Some of these constraints may be of importance in F0 modelling for speech used in man-machine interactions, particularly over the telephone. Over the coming four years of the COST 258 Action, we will be exploring means and the limits of extending the model to a greater variety of intonational contours.

4. STATISTICAL MODELLING OF TIMING

Research on speech timing has documented influences at three levels: the segmental, the syllabic and the phrase level. O'Shaughnessy' influential early model for French was based exclusively on segmental influences (O'Shaughnessy, 1981, 1984). He proposed a rule-based model on the basis of numerous readings of a short text containing all phonemes of French. 33 rules accounted for the modification of a mean segment duration according to segment type, segment position and phoneme context. For sound classes that did not involve pre-pausal lengthening, the model was able to predict durations with a standard deviation of 9 ms, but was less accurate for pre-pausal vowel durations. Moreover, the model was not able to predict silent inter-lexical pauses.

Syllable-sized durations are generally less variable than subsyllabic durations, and thus may represent more reliable anchor points for the calculation of a general timing structure than segmental durations (Barbosa 1994; Keller, 1993; Zellner, 1994). Furthermore, stress variations and variations of speech rate tend to modify at least syllable-sized units. This is taken into account by Bartkova's model (1985) which adds a syllabic coefficient to those for the segment. It depends on the nature of the word (lexical/grammatical), and on word position (initial, medial, final syllable). There is also an "accentuation coefficient" which depends on the next consonant, the presence/absence of a syntactic boundary in the case of a final vowel, or on the presence/absence of clusters in the case of a final consonant, as well as on the syllabic structure near a pause.

According to Bartkova, a comparison of predicted and measured durations in 10 sentences results in a mean difference on segmental duration of ±15 ms. This represents about 1/2 of an overall standard deviation, taken over the totality of our own measures of French segment duration. At first blush, this would appear to be a satisfactory value. However, to be able to judge the quality of the predictive model more adequately, the precision for time-critical consonants (such as plosives) should be examined, and that with respect to specific positions of the segment within the word or the prosodic phrase.

Furthermore, a predictive model needs to incorporate information concerning conditions within the word or the phrase. For example, the prediction of pauses for slow speech requires phrasal knowledge, which is not captured at the segmental or at the syllabic level. In the area of word group boundaries in French speech, a great deal of work has been accomplished to determine the nature of these groups — syntactic groups, prosodic groups, rhythmic groups, intonational groups, the congruence between these labels — and to calculate the automatic generation of such groups and potential inter-group pauses (Delais, 1994; Monnin & Grosjean, 1993; Pasdeloup, 1988).

In our own approach to timing, we attempted to build a robust, basic temporal structure with the smallest number of segmental and syllabic factors. Through a large number of stepwise regressions, parameters were chosen at each succeeding level so as to explain the greatest proportion of the variance in the residue of the previous analysis. In this manner, a three-tier model, based successively on segmental, syllabic and phrasal information, was constructed (Keller & Zellner, 1996). We based our analyses on manually measured segment and syllable durations in a single-speaker corpus of relatively fast French speech. The following parameters were identified as most relevant to segmental duration: (a) the temporal class of the segment preceding the current segment, (b) the temporal class of the current segment, (c) the temporal class of the subsequent segment, (d) the temporal class of the ulterior segment, (e) the position in the prosodic group of the syllable containing the current segment, (f) the grammatical status of the word containing the current segment, and (g) the number of segments in the syllable containing the current segment. "Temporal class" refers to one of nine clusters of typical durations for segmental duration.

A general linear model was calculated for the 2266 first segments of our corpus (Split 1 data), on the basis of these factors. This model explained 51.4% of the variance, i.e., it correlated at r=.717 with its own data set (Figure 4). Coefficients were then derived from the model and were used to predict a Split 2 data set of a second 2266 segments on the basis of factor information for each of its segments. The model explained 48.0% of this new data set, i.e., the values predicted on the basis of coefficients derived from Split 1 data correlated at r=.693 with the measured segmental durations in the Split 2 data set (Figure 5).

Computationally, the calculation of segmental duration precedes Fujisaki modelling, in the sense that the Fujisaki F0 model takes the output of the timing algorithm as its temporal input. The combined effect leaves a more or less

satisfactory auditory impression (see demonstrations at the Workshop). However, improvements in timing are definitely possible, and could possibly be obtained by a more refined quantitative model. But that presupposes a much larger data base than that which we used for the current study.

5. FINALLY: WHAT KIND OF PROSODY?

We have been proceeding on the assumption that a close approximation of human prosody will give us better prosody in speech synthesis. Basically, this seems to be the case. However, the essential difficulty is that there is not *one* appropriate prosodic pattern for each given sentence. Not only are there different patterns for different types of speech, different conditions of attendant noise impediment, but also different degrees of emotional charge, and a large choice of individual personal speech patterns. Choosing and modelling the appropriate type of speech and voice pattern will be as much of a challenge of the coming 15 years as was the development of a solid initial model during the past 15 years. The perfectly soothing, situation-appropriate and corporation-approved voice of HAL in the film "2001" is still some distance in the future.

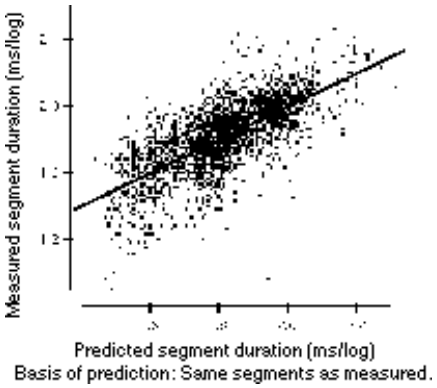

Figure 4. The prediction of segmental duration (x-axis) on the basis of the temporal class of the preceding, current, subsequent and ulterior segment, plus the position of containing syllable in the prosodic phrase, the number of segments in the containing syllable and the grammatical status of the containing word. Here, predictions are compared to the same data from which the model is derived (y-axis, split 1 data, r=.717, N_1=2266).

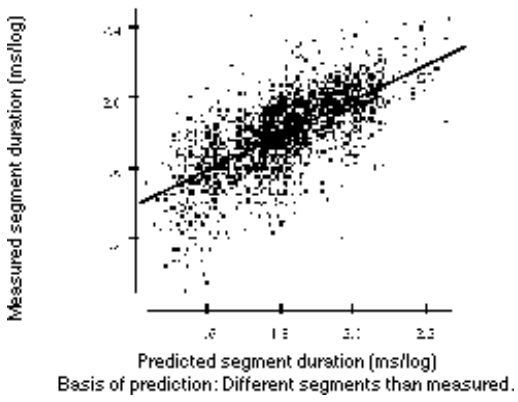

Figure 5. The prediction of segmental duration as in Figure 4, but here, predictions derived from coefficients based on the split 1 data are compared to measures derived from a different data set (split 2 data, r=.693, N_2=2266).

REFERENCES

't Hart, J. T., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge: CUP.

Astesano, C., Di Cristo, A. & Hirst, D.J. (1995). Discourse based empirical evidence for a multi-class accent system in French.

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.

Bartkova, K. (1985). Nouvelle approche dans le modèle de prédiction de la durée segmentale. *14ème Journées d'Etudes sur la Parole.* (pp. 188-191). Paris.

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.

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

Delais, E. (1994). Prédiction de la variabilité dans la distribution des accents et les découpages prosodiques en français. *XXèmes Journées d'Etude sur la Parole* (pp379-384). Trégastel.

Di Cristo, A. & Hirst, D. (1994). Rythme syllabique, rythme mélodique et représentation hiérarchique de la prosodie du français. *Travaux de l'Institut de Phonétique d'Aix, 15.* 13-24.

Fujisaki, H., & Hirose, K. (1982). Modelling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation. In *Preprints of the Working Group on Intonation, 13th Intl. Congress of Linguists* (pp. 57-70). Tokyo.

Keller, E. (1993). *Prosodic Processing for TTS Systems: Durational Prediction in English Suprasegmentals.* Final Report, Fellowship, British Telecom.

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

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.

Pasdeloup, V. (1988). Analyse temporelle et perceptive de la structuration rythmique d'un énoncé oral. *Travaux de l'Institut de Phonétique d'Aix, 11*, 203-240.

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

Pierrehumbert, J. (1981). Synthesizing intonation. *Journal of the Acoustical Society of America, 70*, 985-995.

Pierrehumbert, J. B. (1980). The phonology and phonetics of English intonation. *PhD thesis, MIT [distr. by the Indiana University Linguistics Club]*.

Traber, C. (1991) F0 generation with a data base of natural F0 patterns and with a neural network. *Proceedings of Eurospeech 1991*, 141-144.

Traber, C. (1992) F0 generation with a data base of natural F0 patterns and with a neural network. In: Bailey, Benoit & Sawallis (eds.): *Talking Machines: Theories, Models, and Designs (pp. 287-304).* Elsevier.

Vaissière, J. (1991). Rhythm, accentuation and final lengthening, in J. Sundberg L. Nord, R. Carlson (Eds).*French in Music, Language, Speech and Brain* (pp.108-120). Wenner-Gren International Symposium Series Macmillan Press, Vol. 59.

Werner, S. (1995). Use of a neural network for parameter optimization in Fujisaki models of intonation. Poster presented at *NODALIDA 95 (Nordic Conference for Computational Linguistics)*, Helsinki, July 29.

Zellner, B. (1994). Pauses and the temporal structure of speech. In E. Keller (Ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State-of-the-Art and Future Challenges* (pp. 41-62). Chichester, UK: John Wiley.

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