A STATISTICAL TIMING MODEL FOR FRENCH

E. Keller and B. Zellner

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

ABSTRACT

Numerous factors influence speech timing. Statistical analysis can identify an order of importance and mutual influences between such factors. A three-tiered (segment-syllable-phrase) model was created by a modified step-wise statistical procedure. It predicts the temporal structure of French of a single, highly fluent speaker at a fast speech rate. The model's predictions correlated with the 1204 syllables of the original corpus at $r = 0.846$.

INTRODUCTION

Research on French speech timing has documented influences at the segmental, syllabic and phrase levels. On the basis of numerous readings of a phonetically balanced short text, O'Shaughnessy [1, 2] proposed a model using 33 rules for the modification of segment duration according to segment type, segment position and phoneme context. For sound classes without prepausal lengthening, the model predicted durations with a standard deviation of 9 ms, yet was less accurate for the prediction of prepausal vowel durations.

The model supposes that timing phenomena can be captured by the segment. However, syllable-sized durations are generally less variable than subsyllabic durations, and may thus represent more reliable anchor points for the calculation of a general timing structure [3, 4]. Furthermore, stress variations and variations of speech rate tend to modify at least syllable-sized units, the syllable may be a psycholinguistic perception unit, and it may also be a minimal unit of rhythm. Syllabic duration can be influenced by the position in the prosodic group, the position in the word, degree of stress, the length of the prosodic group, the position according to the stressed syllable, semantic focus, proximity of syntactic boundaries, the lexical or grammatical status of the word, and emotional factors [5-24]. Some of these may be redundant, e.g., lexeme-final position may be redundant with phrase-final position.

Bartkova [5, 6] added supra-segmental coefficients to her formula for segment durations. Some depended on lexical/ grammatical status and on intra-word position, while others depended on the following consonant, the presence of a syntactic boundary, the presence of clusters, or the syllabic structure near a pause. A comparison of predicted and measured durations in 10 sentences gave a mean difference on segmental duration of ±15 ms. Such a difference can be a handicap for short segments. In our corpus the mean duration for /d/ was 50 ms and a 15-30 ms divergence would correspond to a 30-60% error.

The strategy of this study was to issue from segmental predictions, and to treat syllabic information as additional information. Beyond the syllabic level, word- and phrase-level information was also considered (syntactic, prosodic, rhythmic, intonational groups) [8, 15, 17, 19, 20, 25, 26], in order to account for syllable duration with the smallest number of factors. At each succeeding level, relevant parameters were chosen to explain the greatest proportion of the variance in the residue of the previous analysis. In this manner, a three-tier model based on segmental, syllabic and phrasal information was constructed.

METHOD

The Corpus and Segmentation

A fluent speaker of French was recorded with 100 phonetically balanced sentences. He spoke quite rapidly (6.5 syllables/sec. or more), with a normal, unexaggerated intonation. Acoustic recordings were made in studio conditions on DAT-tape. The digitized

data was transferred to computer and was downsampled to 16 kHz.

The time occupied by phonetic segments was labelled with the Signalyze™ program according to a method defined in our laboratory. Specifically, segment transitions were analyzed according to three articulatory levels: labial, lingual and laryngeal. For example, the coarticulatory overlap at the /e/-/s/ transition was marked by symbols representing "onset of frication, associated with the lingual level", followed by "offset of fundamental frequency, associated with a cessation of vocal cord activity". Segmentation reliability was assessed by examining how and where points of transition between inferred articulatory states were marked. Interjudgmental agreement on *robustness* (the application of criteria to state transitions) was scored 1 (low) to 3 (excellent), and agreement on *precision* was scored on 1 (more than two Fo periods difference) to 3 (less than 1 Fo period difference in measurement). Over 50 types of state transitions, there were no cases of low robustness or low precision. The average robustness was 2.53 and the average precision was 2.68.

Analysis and Results

A modified step-wise statistical regression technique for segmental, syllabic and phrase level information was used to develop a model of the speaker's timing behaviour. An issue concerned the calculation of segment duration in a corpus where coarticulatory transition zones are marked explicitly. Is segment duration considered to be the steady-state portion of the signal, or does it include one or both zones of acoustically prominent coarticulatory overlap with adjoining segments? The issue was resolved with reference to durational variation. Since the coefficient of variation over the three zones was systematically smaller (average 0.375) than that of the steady-state zone (average 0.412), the combined duration of the three zones was considered to correspond to "segment duration". Syllable durations were constructed from segment durations by taking into account transitional overlaps (i.e., syllable 2 was overlapped with syllable 1).

The Segmental Model

Raw segment durations were non-normal in their distribution and a log transformation produced a close approximation to a normal distribution. Subsequent to log transformation, segments were grouped according to their mean durations and their articulatory definitions. Eight types of segments could thus be identified. Groups showed roughly comparable coefficients of variation, and an inspection of histograms and normal probability plots showed roughly normal distributions for all classes whose N was greater than 100.

Using the Data Desk® statistical package, a general linear model for discontinuous data (based on an ANOVA) was calculated with partial sums of squares. The following main and interaction factors (up to two-way) were postulated: Duration ($\log_{10}$(ms)) = constant + previous type + current type + next type + previous type * current type + current type * next type + previous type * next type.

Expressed in terms of a Pearson product-moment correlation, the model's predicted segmental durations correlated with empirical segment durations at $r = 0.696$. To test Model 1 in the syllabic context, syllable durations were calculated and were compared to measured syllable durations. The correlation coefficient was $r = .647$ ($N = 1203$, $p < .0001$). The residue from the model (= observed - predicted) was termed "Delta 1" and served as the basis for further factorial modelling at the syllabic level.

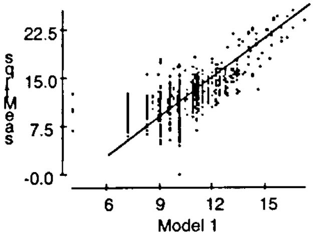

Figure 1. Prediction of the Segmental Model (Model 1): Syllable durations predicted by segmental durations ($r = .647$). Values of Figures 1-3 are in sqrt(ms).

The Syllabic Model

After much experimentation with syllable-level factors described in the literature, a three-factor model, including two-way interactions, was retained for the syllabic analysis: delta 1 = constant + function + position + schwa + function * position + function * schwa + position * schwa, where *"function"* distinguishes lexical vs. function word status, *"position"* identifies three positions in the word, "monosyllabic and polysyllabic-initial", "polysyllabic pre-schwa" and "other", and "schwa" indicates whether or not a schwa is present in the syllable. All main and interaction factors were significant at $p<.05$ by ANOVA.

Syllable durations obtained from the segmental model were additively combined with those for Delta 1 to produce the Syllabic Model (Model 2). Syllable durations showed roughly a square root distribution and were square-root transformed before analysis. Predictions for syllable durations were correlated with transformed observed durations at r = .723 (N=1203) (Figure 2). The residual data from this model was termed Delta 2.

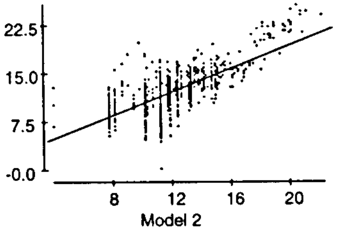

Figure 2. Prediction of the Syllabic Model (Model 2): Syllable durations predicted by segmental durations and syllable-level factors (r = .723).

The Phrase Model

Predictions of Models 1 and 2 showed a noticeable deviation from the regression line in the higher values. Specifically, most syllable durations in the > 280 ms range were underestimated. Furthermore, Delta 2 showed the most pronounced residual error for utterance-final syllables ending in a consonant. A phrase-final correction term was thus calculated for Model 3.

The predictions of Model 3 correlated with the observed square root-transformed syllable durations at r = .846 (Figure 3). The residual values from Model 3 varied quasi-randomly around 0. At the present time, it appears that only more sophisticated rules for the generation of the schwa vowel may still be able to improve this model's predictive capacity to some degree.

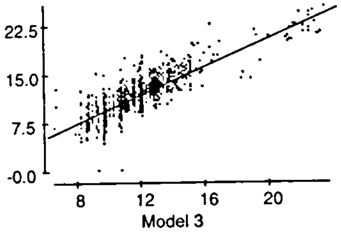

Figure 3. Prediction of the Phrase Model (Model 3): Syllable durations predicted by segmental durations, syllable-level factors and phrase-final lengthening (r = .846).

DISCUSSION

A general model for the prediction of the fast-speech performance of a highly fluent speaker of French was constructed. In view of current discussions surrounding segmental and syllabic contributions to timing models, it is interesting to note that segmental information accounts for a major portion of the variance explained by the model.

The correlation of 0.846 between predictions of Model 3 and the original data set is encouraging. Further improvements in the modelling may come about by the prediction of the presence vs. the absence of schwa, by explicit prediction of speech rate manipulation, and in longer texts, by a better modelling of pauses.

In the present fast-speech corpus, no phrase-level effects other than phrase-final lengthening were identified, in contrast to our findings on the production of French at a normal speech rate, where a systematic increase of lexeme-final syllable durations was observed over the extent of the prosodic phrase [25]. It seems likely that in conditions of considerably accelerated speech rate, our speaker sacrificed some of the "niceties" of phrase-internal timing modulation, and limited himself to a single, phrase-final durational marker.

ACKNOWLEDGEMENTS

Acknowlegements to N. Thévoz, A. Enkerli, H. Mesot, C. Bourquart, N. Blanchoud, and T. Styger. The research is supported by the Fonds National de Recherches Suisses (Projet Prioritaire en informatique, ESPRIT Speech Maps) and by the Office Fédéral pour l'Education et la Science (COST-233).

REFERENCES

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

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

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

[4] 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.

[5] Bartkova, K. (1985). Nouvelle approche dans le modèle de prédiction de la durée segmentale. *14ème JEP* (pp188-191). Paris.

[6] Bartkova, K. (1991). Speaking rate in French application to speech synthesis. *XIIème Congrès International des Sciences Phonétiques*, (pp 482-485). *Aix en Provence. Actes.*

[7] Campbell, W.N. (1992). Syllable-based segmental duration. *Talking Machines. Theories, Models, and Designs* (pp. 211-224). Elsevier Science Publishers.

[8] 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* (pp. 379-384). Trégastel.

[9] Duez, D., Nishinuma, Y. (1985). Le rythme en français. *Travaux de l'Institut de Phonétique d'Aix, 10*, 151-169

[10] Duez, D. & Nishinuma, Y. (1987). Vitesse d'élocution et durée des syllabes et de leurs constituants en français parlé. *Travaux de l'Institut de Phonétique d'Aix, 11*, 157-180.

[11] Fant, G., Kruckenberg, A., Nord, L. (1991). Durational correlates of stress in Swedish, French and English. *Journal of Phonetics. 19*, 351-365.

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

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

[14] Grosjean, F. (1983). How long is the sentence? Prediction and prosody in the on-line processing of language. *Linguistics, 21*. 501-529.

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

[16] Konopczynski, G. (1986). Vers un modèle développemental du rythme français: Problèmes d'isochronie reconsidérés à la lumière des données de l'acquisition du langage. *Bulletin de l'Institut de Phonétique de Grenoble, 15*, 157-190.

[17] Martin, Ph. (1987). Structure rythmique de la phrase française. Statut théorique et données expérimentales. *Proceedings des 16e JEP* (pp. 255-257). *Hammamet.*

[18] Mertens, Piet. (1987). *L'intonation du français. De la description linguistique à la reconnaissance automatique.* Thèse doctorale, Katholieke Universiteit Leuven.

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

[20] 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.

[21] Pasdeloup, V. (1990). *Organisation de l'énoncé en phases temporelles: Analyse d'un corpus de phrases réitérées*, (pp. 254 - 258). 18émes Journées d'Etudes sur la Parole. Montréal, 28 - 31 Mai.

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

[23] Wenk, B. J. & Wiolland, F. (1982). Is French really syllable-timed? *Journal of Phonetics, 10*, 177-193.

[24] Wunderli, P. (1987). *L'intonation des séquences extraposées en français.* Tübingen: Narr, 1987.

[25] Keller, E., Zellner, B., Werner, S., & Blanchoud, N. (1993). The Prediction of Prosodic Timing: Rules for Final Syllable Lengthening in French. *Proceedings*, ESCA Workshop on Prosody (pp. 212-215). Lund, Sweden.

[26] Saint-Bonnet, M. & Boë, J. (1977). Les pauses et les groupes rythmiques: leur durée et disribution en fonction de la vitesse d'élocution. *VIIèmes Journées d'Etude sur la Parole*, (pp. 337-343). Aix en Provence.