# Intensity Prediction for Speech Synthesis in French

K. Bartkova - P. Haffner - D. Larreur
France Télécom - CNET Lannion A - TSS/RCP
Route de Trégastel - 22300 Lannion - France

## ABSTRACT

*The goal of the present study is to predict sound intensity for speech synthesis in French. In order to set up a model for intensity prediction, we first studied intensity variation in natural speech. The data base used was hand segmented and phonetically and syntactically labelled. The results of this part of the study were introduced into a rule-based model whose parameters were subsequently optimised using stochastic gradient procedure. Next, a neural network based model was developed and trained using part of the labelled data.*

## INTRODUCTION

Sound intensity is considered to be the least important of the three prosodic parameters for the perception of synthetic speech quality. Most of the time, researchers have settled for a decrease in sound intensity at the end of the sentence (Calliope 1989) while maintaining fairly constant intensity elsewhere. As far as diphone speech synthesis is concerned, the intensity of the basic units stored in the dictionary has been kept unchanged during speech generation. Some studies (Granström 1991), however, have highlighted the role of intensity in speech synthesis for modelling different styles of speech.

## INTENSITY VARIATION IN NATURAL SPEECH

In order to set up a model for sound intensity prediction in French, we first investigated intensity variation in a corpus of about 1 hour of natural speech recorded by a male speaker. This corpus contained isolated sentences as well as 9 short texts. Its structures covered the majority of linguistic and prosodic possibilities in French. The intensity of the whole corpus was normalised (sentence by sentence) to its highest value. Intensity values were measured in dB in approximately the middle of the each phoneme thus avoiding accidental humps (if any) known to sometimes occur in unvoiced fricatives.

Even if the corpus used didn't contain sentences dedicated to the intrinsic intensity study, it afterwards became possible to bring out this value. Indeed, as the corpus was large enough, constraints could be imposed when sounds where chosen for intrinsic intensity calculation. Duration, pitch value and the left and right context of each sound were thus controlled. This way, for example, only vowels with a duration between 100 and 150 ms, pitch value between 70-120 Hz and left and right context belonging to the same consonantal classes were used for calculating intrinsic intensity. Since large variations can exist from one group to another[1], duration threshold was determined for each consonantal group. A distinction was made between the two allophones of /R/ : one was voiced (surrounded by voiced contexts) and the other voiceless (surrounded by unvoiced contexts). For stop consonants, voiceless stop intensity was measured during their burst whereas voiced stop intensity was measured during stop closure.

[1] The mean duration of unvoiced fricatives is much longer than that of semi-vowels.

Our findings on the intrinsic values of vowel and consonant intensity (Fig.1 & 2) are very similar to those observed in other studies (Di Cristo 1978).

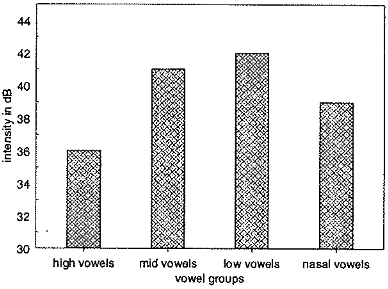

**Figure 1.** *Intrinsic intensity for vowels.*

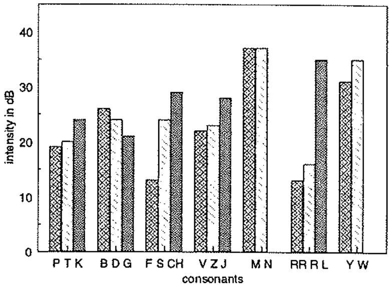

**Figure 2.** *Intrinsic intensity for consonants*[2].

The first part of this study revealed that duration, pitch, left and right contexts and sound position (with respect to the word and pause occurences) are among the most important parameters for detemining sound intensity. When the influence of one of these parameters on sound intensity was studied (duration, pitch, contexte...), the others were only allowed to vary between controlled thresholds. In this way, we hoped to clarify the relationship between sound intensity and other phonetic or syntactic events.

The number of sentences containing word focus was quite small in our corpus and therefore it would be dangerous to draw general conclusions. Nevertheless, it can be noted that increased sound intensity was observed in focus syllables.

## RULE-BASED MODEL

The findings from the first part of our study were introduced into the rule-based model under different rules. The prediction formula took into account intrinsic sound intensity along with some additional coefficients expressing the influence of the previously discussed phonetic and syntactic parameters. The following formula was used for intensity prediction :

$$\textit{Sound\_Intensity} = \textit{intrinsic\_int.} + \textit{coef}_{(\textit{left\_context})} + \textit{coef}_{(\textit{right\_context})} + \textit{coef}_{(\textit{duration})} + \textit{coef}_{(\textit{pitch})} + \textit{coef}_{(\textit{position})}.$$

An intrinsic intensity value was determined for each phoneme. The parameter coefficients relevant to the model (context, duration...) could be positive or negative. They were determined for a homogenous group of sound (ex: unvoiced plosives, voiced plosives, nasals, semi-vowels...).

Mean error prediction was calculated as the mean value between the predicted value and the target value (measured value). Figures 3 and 4 represent the histograms of prediction dispersion for vowels and consonants. The ***intensity prediction error rate*** for the ***whole corpus*** was ***4.3 dB for vowels and 4.9 dB for consonants (leading to an average error rate of 4.6 dB per sound).***

Coefficient and intrinsic intensity values were subsequently optimised using stochastic gradient procedure. The corpus was split into two parts. One part was used for "tuning" model values; the other for testing new values (data was split as in the NN-model). The prediction error rate was reduced by about 1 dB per sound (i.e. 20% reduction).

[2] **"J"** was used for the French fricative /ʒ/ as in "jour"; **"RR"** was used for the unvoiced /R/ allophone; **"Y"** was used for the semi-vowel /j/ as in "bien".

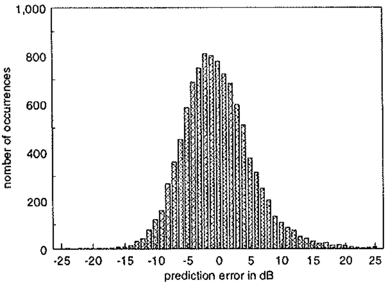

**Figure 3.** *Vowel prediction dispersion.*

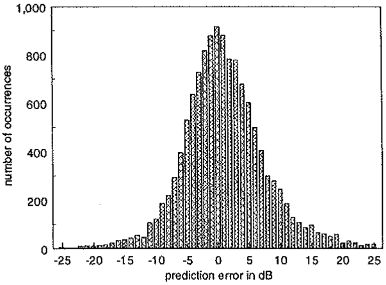

**Figure 4.** *Consonant prediction dispersion*

**Table 1.** *Intensity prediction error in dB for vowels and consonants provided by the rule-based model before and after optimisation by a stochastic gradient technique.*

| | vowels | | consonants | | all sounds | |
|---|---|---|---|---|---|---|
| | **training** | **testing** | **training** | **testing** | **training** | **testing** |
| **rule** | **4.6** | **4.6** | **4.9** | **5.0** | **4.8** | **4.8** |
| **stoch. gr.** | **3.3** | **3.4** | **4.0** | **4.2** | **3.7** | **3.8** |

## NEURAL NETWORK MODEL

The Neural Network (NN) used in this study was developed with the Aspirin/Migraine software. A three layer NN was created in order to approximate the function:

***Intensity_Code = f(Phoneme_Code, Left_cont_Code, Right_cont_Code, Position_Code, Pitch_Code, Duration_Code).***

Phonetic knowledge was used to set up numeric codes for the relevant parameters, which were afterwards normalised to facilitate the training procedure convergence. The network architecture consisted of an input, a hidden and an output layer. The input layer gathered the cells coding for the input parameters cited above. The hidden layer was completely connected with the input layer. The output layer contained 36 cells which were arranged in the form of a thermometer. The maximum intensity was 36 dB with 18 above and below the mean value. Because of the considerable difference in vowel and consonant intensity, two separate networks were trained to predict them.

The NN was trained with the Gradient Back-Propagation procedure. Between 5 to 40 hidden cells were tested in the training procedure.

The ***mean error rates*** for ***vowel and consonant intensity*** prediction were ***3.6 dB and 4.5 dB respectively (on average, about 4 dB per phoneme)***. The number of hidden cells had little effect on results. The prediction error rate for consonants was the mean value provided by 8 networks trained for different groups of consonants.

## PRELIMINARY AUDITORY TESTS USING NATURAL SPEECH

Phoneme intensity of 40 isolated sentences (belonging to our testing data) was modified according to the model parameters[3]. 20 listeners were asked to judge the quality of these sentences and in so doing compare altered and natural sentence intensity. The paired sentences (natural A and model-provided B) were presented in random order. For each

[3] All the other prosodic parameters (sound duration and pitch) were kept unchanged.

sentence pair listeners noted whether or not they had a preference for one of them. The following results were obtained :

**Table 2.** *Auditory test results in % for rule-based model and NN model.*

| | no preference | pref. for nat. int. | pref. for mod. int. |
|---|---|---|---|
| **rule-based mod.** | **46.3** | **28.3** | **24.2** |
| **NN model** | **31.9** | **40.41** | **27.6** |

Thus in **70.5%** of the time, the rule-based model provided intensity and in **59.5%** of the time, the NN model provided intensity was considered by listeners as fully natural.

We are well aware that natural signal can more readily undergo modifications without quality corruption than synthetic signal. Nevertheless, results obtained by the test are encouraging and it can be hoped that no major problems will be encountered when the intensity model is implemented to speech synthesis.

## IMPLEMENTATION TO SPEECH SYNTHESIS

In this part of sudy we had to cope with the problem of sound level differences between the diphones and sounds of our corpus. This is why the absolute intensity value provided by the model was converted into a multiplicative coefficient expressing the slope movement of sentence intensity variation. As great care was taken to record diphones with a cosntant intensity; thus it was possible to consider diphone intensity as an intrinsic intensity value. The diphone intensity is then subject to modification depending on event at the sentence and sound level.

## CONCLUSION

The aim of this study was to set up a model for sound intensity prediction for speech synthesis in French. We hope that controlling all three of the prosodic parameters will increase the resemblance of synthetic to natural speech. Accurate energy prediction at the sentence level will improve the perception of speech fluency by eliminating unpleasant, too salient sounds which occur in unexpected positions. Last but not least, controlling sound intensity will help to introduce the perception of sound depth into speech synthesis. It is true that sentence focus can be satisfactorily modelled by appropriate pitch movement and sound duration. But adding intensity will improve naturalness and the impression of text comprehension by the reading system.

## ACKNOWLEDGEMENTS

The authors would like to thank F. Agashira and B. Cherbonnel, for implementing the signal processing part of intensity control, and for their valuable comments and great patience.

## REFERENCES

Calliope (1989) *"La parole et son traitement automatique"*, Masson.

A. Di Cristo (1978) *"De la microprosodie à l'intonosyntaxe"*, Thèse d'Etat, Université de Provence, Aix-en-Provence.

B. Granström, & L. Nord (1991) "Neglected dimensions in speech synthesis", *Proc. of the ESCA workshop,* pp. 27-1 - 27-5; Barcelona, Spain.