USING SEGMENTAL DURATION PREDICTION FOR RESCORING THE N-BEST SOLUTION IN SPEECH RECOGNITION
K. Barkova, D. Jouvet & T. Moudenc
CNET, France Télécom, Lannion, France

ABSTRACT
The aim of this study is to set up a rule-based model of segmental duration in French for an automatic speech recognition system. This model was introduced at the post-processing stage of speech recognition in order to rescore the N-best solution hypotheses. In preliminary experiments conducted on isolated and connected word databases, the reduction in the recognition error rate introduced at the post-processing stage prosodic parameters such as F0, sound intensity.

INTRODUCTION
Prosodic features contain valuable information about speech structuring. A great number of studies have focused their attention on the measure and description of principal physical prosodic parameters such as F0, sound duration and sound intensity. Researchers in automatic speech processing have long been aware of the importance of integrating prosody into different automatic systems. In speech synthesis the role of prosody is clearer: it must be modelled for generating natural sounding speech. In automatic speech recognition, prosodic parameters have primarily been used in order to segment signals into prosodic units. The main prosodic cues used for signal segmentation are: final syllable lengthening of a prosodic constituent, and F0 movement amplitude.

A great number of studies deal with sound duration modelling in recognition systems based on Hidden Markov Modelling (HMM). Some of these studies use minimal sound duration or sound duration normalized by utterance length, or variable sound duration according to speech rate.

THE RULE-BASED MODEL
The aim of this study is to introduce phonetic knowledge into sound duration modelling in order to set up a phonetically-based sound duration model. This phonetic duration model is then used in the post-processing of the N-best solutions hypotheses given by an HMM based system using only spectral representation information. The role of the duration prediction is to distinguish between good and bad solutions proposed by the recognizer. Thus, either the duration score confirms the scoring of the HMM, or, conversely, penalizes the score (for example, when the duration of the segments constituting the solution does not match the duration predicted by the model).

The rule-based duration model for French sounds predicts segmental duration according to relevant phonetic and phonologic events. Phonemes are grouped into macro-classes. There are 4 vocalic macro classes (oral vowels, nasal vowels, neutral schwa-like vowels and semi-vowels), and 7 consonantal macro classes (voiceless plosives, voiced plosives, voiceless fricatives, voiced fricatives, nasals, r and l). For each macro class, mean phoneme durations and standard deviations were calculated according to: the left and right context (also expressed in macro classes), the word length, and the position of the syllable in the word (final syllable versus non-final syllable).

CONTEXT GROUPING
Several studies in micro-prosody illustrate that right consonant contexts have greater influence on vowel duration than left consonant contexts. In French, the accent falls on the last syllable of a prosodic unit. As such, the vowel duration is clearly dependent upon the right context only in final "stressed" syllables. In order to obtain appropriate phonetic modelling, considering the right context of the last syllable of a lexical word is sufficient. Unfortunately, HMM segmentation is not always accurate. A type of "spectral inertia" persists in the segmentation process. This is due to the fact that parameters used for modelling contain mainly spectral information (8 Mel Frequency Cepstral coefficients and their first and second order temporal derivatives), and only three values related to energy (energy value and its first and second order derivatives). One can assume that mistakes made by HMM in segmentation are consistent for this very reason. In order to surmount this segmentation defect, both left and right contexts are taken into account in the sound duration modelling. In terms of syllable vowel duration, when a consonantal cluster closes the syllable, in addition to the immediate right consonantal context, the last consonant is also taken into account. This is because the phonetic characteristics of the last consonant (together with syllable structure) influence vowel duration.

PARAMETER SMOOTHING
Smoothing was implemented when the number of occurrences was not high enough to enable the reliable estimation of a sound duration parameter. Among the different phonetic parameters an a priori hierarchy was established, which specifies the order in which the smoothing is conducted. During rule smoothing, all the other parameters remain unchanged. For example, for a vowel, the first phonetic parameter to be considered is the left context, while the other conditions (right context, syllable position, word length) stay unchanged.

MODEL PARAMETERS
Two different models were set up using the three corpora described herein. Two training procedures were evaluated: one was corpus-dependent, and one was pluri-corpus. Corpus-dependent means that the sound duration parameters were trained and used on the same corpus.
Sound Duration Prediction Errors

Figure 2: Standard Deviation of the Sound Duration Prediction Errors

Table I: Percentage of Correct Identification using only Duration Information

<table>
<thead>
<tr>
<th></th>
<th>Digits</th>
<th>Numbers</th>
<th>Trégor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration prediction error</td>
<td>45 %</td>
<td>62 %</td>
<td>64 %</td>
</tr>
<tr>
<td>Speaking rate</td>
<td>68 %</td>
<td>38 %</td>
<td>51 %</td>
</tr>
<tr>
<td>Duration prediction error + Speaking rate</td>
<td>76 %</td>
<td>67 %</td>
<td>80 %</td>
</tr>
</tbody>
</table>

Figure 3: Speaking Rate Coefficients

MODEL EVALUATION

The efficiency of the sound duration predictive model was evaluated in the post-processing procedure. The benefits of different components (e.g., prediction error and speech rate coefficients), were initially tested separately, then together, and finally in combination with HMM scores.

Detailed results show that this processing proved to be beneficial, as did the possibility of recovering HMM errors with duration prediction. Table I provides correct recognition percentages using duration and speech rate information, separately and combined. The percentage of the HMM errors recovered by duration post-processing was about 50% for each corpus, duration and speech rate scores combined.

As illustrated in Table I, speech rate modelling works well for short, mainly monosyllabic vocabularies, such as digits. Duration modelling works better for longer word vocabularies (such as Numbers or the Trégor). Regardless, the combination of both scores provides relatively good results, considering that only duration information was used.

Table II illustrates the recognition error rate reduction, following the introduction of the duration prediction error score and/or the speech rate score in post-processing, in comparison with HMM alone. Preliminary tests were conducted, recombining duration information and a supplementary parameter obtained from an a priori segmentation of the speech signal (this parameter expresses the number of stationary zones that occur in each segment of the signal) [6].

CONCLUSION AND DISCUSSION

This study focuses on the investigation of new parameters used in speech recognition system post-processing. It is reasonable to assume that the introduction of different types of parameters can add valuable information to the rescore of the HMM spectral score, where scoring is achieved using speech spectral representation following an initial pass through the system. Sound duration rule-based prediction is one type of supplementary parameter. Although information supplied by sound duration is poorer than those of spectral word representations (two different words or hypotheses can have exactly the same duration), initial attempts at evaluating the efficiency of these parameters have proved quite helpful.

Further attempts to introduce other prosodic parameters such as F0, in order to rescore the N-best solutions in post-processing, are also being made.

REFERENCES