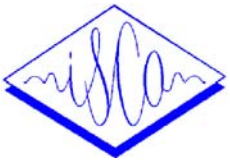

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SELECTIVE PROSODIC POST-PROCESSING FOR IMPROVING RECOGNITION OF FRENCH TELEPHONE NUMBERS

Katarina Bartkova & Denis Jouvet
France Télécom, CNET/DIH/DIPS, 2 Av. Pierre Marzin
22300 Lannion, France
{katarina.bartkova, denis.jouvet}@cnet.francetelecom.fr

10.21437/Eurospeech.1999-70

ABSTRACT

This study describes a selective prosodic post-processing procedure for improving the recognition of telephone numbers in French. The aim of the post-processing procedure is to recover recognition errors made by an HMM based ASR system. Instead of a global post-processing, this paper proposes a selective one. Post-processing is carried out only on some recognised numbers and only if its associated frequent confusion is also present in the N-best candidates. In such a case the discrimination between the solutions is carried out by checking the duration of a specific segment in a pertinent prosodic position. On the different data used in this study, about 23 % of the substitution errors are considered as being possible to recover with the selective duration post-processing and of this amount about 40 % of the errors are actually recovered.

Key-words: speech recognition, post-processing, phone duration.

1. INTRODUCTION

Recognising telephone numbers is an important task in some telecommunication applications. Unlike in English where telephone numbers are pronounced digit by digit, in French 2 or 3 digit numbers are generally used. The required vocabulary is therefore larger in French and lot of errors stem from a few confusions between pairs of numbers. Similar number pairs are uneasy to distinguish especially in difficult environments such as mobile calls. However, such pairs are acoustically very similar and they generally obtain close acoustical scores. Therefore, in case of error, the correct answer is often present among the first few N-best candidates. Consequently, an adequate post-processing should allow recovering part of these errors.

Segmental post-processing have been used for several years. At first it was mainly used for re-scoring the N-best hypothesis [1,2,3]. The formalism used in [2] relies on a double parameter modelling associated to correct and incorrect segments. This formalism implies the computation of likelihood ratios. It was then used to reject out-of-vocabulary data in various tasks using various phonetic and prosodic parameters [4]. In the studies dealing with the post-processing issue, the set of segmental features was applied systematically to every hypothesis and to every phonetic unit of it.

The present study is a pilot one in which we propose an approach based on a selective post-processing in order to improve the recognition of telephone numbers in French. The aim of this study is also to use prosodic parameters, which have proven to be efficient in previous studies and are not used by the HMM-modelling in the first recognition path. The parameter used here to post-process the N-best candidates is the segment (phone or word) duration, although other parameters, such as F0 or voicing features, are also currently under study in order to recover errors which cannot be, or are not well enough recovered, by phone duration alone.

2. METHODOLOGY APPLIED

It appeared interesting to analyse the recognition errors in order to decide which parameter is the most appropriate to double check the N-best candidates delivered by the HMM and which is the pertinent part of the word or sentence where this checking is to be carried out.

After a first error analysis it appeared that a large amount of substitution errors could be recovered by phone duration modelling. In this first approach, therefore, only the phone duration modelling is considered in the error recovering tests. In fact, about 24% of the all substitution errors were post-processed by segment duration.

The difficulty to use duration to post-process recognition hypothesis stems from the fact that many of them have very similar phone or word duration. Moreover, the duration is elastic and allows speakers quite a great variation without altering the meaning of the word. Therefore the major concern of this approach was to determine the pertinent phone position where the difference between the compared duration of the segments is robust enough to resist to inter-speaker variability.

Thus, the duration post-processing is carried out in a selective manner, that is only some of the HMM

solutions are double checked. In order to apply the duration post-processing, two conditions has to be fulfilled. Firstly, the pairs of numbers double-checked by the duration post-processing must be considered as significantly different as far as the duration of the phone occurring on a duration-pertinent position is concerned. Secondly, the first number of the pair must be present in the first (best) decoding and the second number must occur in one of remaining N-best candidates, example:

first solution:	05	61	13	***80***	34
N-solution:	05	61	13	***81***	34

The duration-pertinent positions of the phones are determined by rules governing the phone duration in French [6]. In French, the stress has a fixed position on the last syllable of the prosodic group. In many languages the last syllable of a prosodic group followed by a pause is longer than the same syllable when it is not. In French, this tendency is reinforced by the position of the stress and by its major physical parameter which is the syllable duration: a stressed syllable, which coincides with the last syllable of the prosodic group is systematically longer than a non-stressed syllable. One can therefore consider that a duration of a syllable occurring once in a stressed (final) position of a prosodic group and once in a non-stressed (prosodic-group-internal) position can be a reliable duration indicator.

Most of the time the five numbers, building up a telephone number, are uttered as separate prosodic units often followed by a short pause. Every telephone number is uttered in a sentence-like manner. Prosodic parameters are used to express major continuation movements on the last syllables of the internal numbers and sentence final movement on the last number of the recording. A particular attention is focused here on number-pairs in which a syllable, present in the two numbers, occurs in one word in a stressed position and in the other in a non-stressed position. For example, when a speech signal segment is recognised once as the number 30 [trɑ̃t] (best candidate) and once as the number 31 [trɑ̃teɛ̃] (one of the N-best candidates), the pertinent phone duration is the duration of the vowel [ɑ̃]: it occurs once (in number 30) in a stressed position in the other (in number 31) in a non-stressed position.

As it is illustrated in Figure 1, in spite a possible erroneous HMM segmentation, the difference between the duration of the two phones for the word-pair 30-31 is quite clear-cut. However, if the HMM segmentation was always phonetically consistent, the approach presented here would be extremely efficient. Unfortunately, segmentation errors, committed by automatic alignment, partly corrupts the pertinent duration post-processing

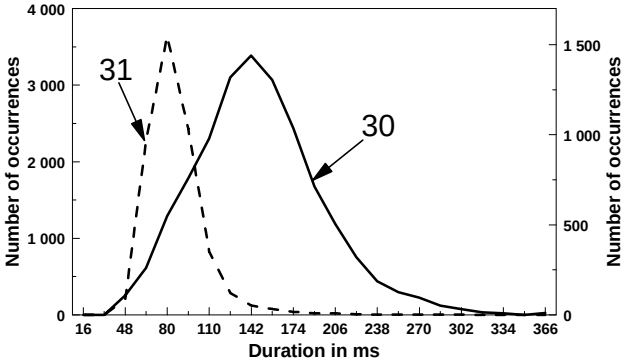

Figure 1: Histogram of the Duration of the Pertinent Vowel [ɑ̃] in the Numbers 30 and 31.

3. CORPUS USED

The corpora used in this study were recorded through the telephone network and contained French telephone numbers. French telephone numbers are made up of 10 digits and are usually pronounced as a sequence of 5 2-digit numbers (i.e. numbers between 00 and 99). However special numbers such as those, starting by 0800 or 0801 are pronounced 0-800 and 0-801 thus implying the recognition of some 3-digit numbers too.

Two signal qualities were recorded, one collected through mobile phones (GSM environment) and the other through fix phones (PSN environment). 17763 telephone numbers were used for the error analysis and the training of the pertinent phone duration. Evaluations were conducted on 17754 telephone numbers from the mobile GSM environment and 22751 telephone numbers from PSN environment.

The automatic speech recognition system used in this study was the PHIL-SOFT HMM-based recognition system [5]. The acoustic modelling units were the allophones and the evaluations were carried out in a speaker independent manner.

4. ERROR ANALYSIS

4.1 Classes of Errors

Recognition errors considered as possible to be recovered by duration post-processing can be grouped into 4 classes:

Group I. A stressed single vowel is recognised as two adjacent vowels. Three number-pairs belong to this group:

80 → 81 [katrœvɛ̃ → katrœvɛ̃ɛ̃]
800 → 801 [yisɑ̃ → yisɑ̃ɛ̃]
80 → 91 [katrœvɛ̃ →katrœvɛ̃õz]

Group II. Poor signal segmentation causes either an omission or an insertion of the final part of the

number. Example of such erroneously recognised numbers-pairs is:

30 ↔ 37 [trãt → trãtsɛt]

Group III. Confusion of the last unvoiced stop consonant, which shortens the preceding adjacent vowel in stressed position, with a voiced fricative consonant, which lengthens such a vowel. An important phenomenon in French is a phonological lengthening of the vowel duration by voiced fricatives and the vibrant [r]. This indicates that besides a language-independent physiological vowel lengthening caused by voiced consonants there is an extra language-dependent lengthening caused by rules governing the phone duration in French. Two number pairs were concerned by this kind of error:

67 ↔ 76 [swasãtsɛt ↔ swasãtsɛz]
87 ↔ 96 [katrœvɛ̃sɛt ↔ katrœvɛ̃sɛz]

Group IV. A word internal unstressed syllable is either omitted or inserted. This kind of error caused confusion between the following numbers:

68↔78 [swasãtʏit ↔ swasãtdizʏit]
69↔79 [swasãtnœf ↔ swasãtdiznœf]
88↔98 [katrœvɛ̃ʏit ↔ katrœvɛ̃dizʏit]
89↔99 [katrœvɛ̃nœf↔ katrœvɛ̃diznœf]

4.2 Error Distribution

Table I. summarises the distribution of the errors according to the four main error groups described in the previous paragraph. The percentage for the different groups (column 2) are obtained as the ratio between the numbers of errors belonging to each group and the total number of errors which can be recovered by duration post-processing. The third column contains the number of the different number-pairs pertinent for the duration post-processing while the last column contains the average number of errors per pertinent number-pair for each group.

Gr.	% of error belonging to the group	number of different number-pairs	average number of errors per number-pair
I	31.0 %	3	193
II	63.7 %	71	17
III	3.3 %	4	21
IV	2.0 %	8	12

Table I: Distribution of the Errors According to the 4 Main Error Groups.

As it appears from Table I, nearly 2/3 of the recognition errors belong to the second group. However, this is the group, which also contains the highest number of different number-pairs (71) what lowers considerably the average number of errors per number-pair (only 17). On the opposite, the first group, which accounts for nearly 1/3 of the errors has a very low number of word-pairs and therefore a very high average number of errors for every number-pair. Thus, it appears extremely interesting to focus on the first error group, as it is the one, which concentrates the highest confusions per number-pair.

5. DURATION MODELING

As already mentioned, for every number-pair a pertinent phone position is determined and only in this position is the phone duration compared. Dealing with short clauses, and considering that most of the time the pertinent position is a stressed one (last prosodic group syllable) it does not appear necessary to use normalised phone duration. As a matter of fact, in French the prosodic group length has little or no influence at all on the duration of the last stressed syllable, especially if it is followed by a pause.

In the first error group (group I) the duration of the two last adjacent vowels in one number is compared to the duration of the last single vowel in the other:

$$\begin{bmatrix} \text{katrœv}\{\tilde{\varepsilon}\} \\ \text{katrœv}\{\tilde{\varepsilon}\,\tilde{\varepsilon}\} \end{bmatrix}$$

In the groups II and III the pertinent duration considered is the last common vowel, different from the neutral schwa like vowel, present in both words. Thus, for example, the comparison between the pair 30 ↔ 32 is done on the duration of the vowel [ɑ̃]:

$$\begin{bmatrix} \text{tr}\{\tilde{\alpha}\}\text{t(ə)} \\ \text{tr}\{\tilde{\alpha}\}\text{t(ə)dø} \end{bmatrix}$$

In the group IV, the word-pair comparison is carried out according to the whole word duration.

For each word-pair the duration of the pertinent segment is compared to a threshold. The thresholds are determined from histograms and accept most of the correct data. Therefore, when, for instance, a phone duration in a stressed position is shorter than the chosen threshold, the corresponding word is considered as misrecognised and then its duration counterpart (second word of the pair), if present in one of the N-best solutions, is proposed as the correct recognition.

6. EXPERIMENTS

Evaluations were carried out on two data types, on mobile phone speech data (generally more noisy) and on fix phone speech data. The mobile phone

corpus (GSM environment) was split into two parts, one part was used for error analysis and for duration training and the other part was used for testing the phone duration post-processing efficiency. The entire fix telephone corpus (PSN environment) was used only for testing.

Table II reports the results obtained on the PSN telephone corpus and on the training and testing mobile corpora. The table contains the percentage of the total number of errors, which could be post-processed by phone duration (second column). The third column contains the percentage of errors actually post-processed by duration (the second element of the number-pair is present in one of the N-best solutions). The last column contains the percentage of the errors recovered by the post-processing (the number of errors added by the duration post-processing is subtracted from the number of actually recovered errors).

corpus	**% of error to process**	**% of error processed**	**% of errors recovered**
Mobile (Training)	24.2 %	74.7 %	42.4 %
Mobile (Test)	24.2 %	76.6 %	36.6 %
Fix (Test)	22.1 %	79.5 %	46.6 %

Table II: Possible and Actual Error Recovery Rates.

The use of the whole word duration to compare number-pairs (group IV) turned out to be completely inefficient and was therefore abandoned. A thorough analysis of errors showed differences in error recovering according to the four error groups defined before.

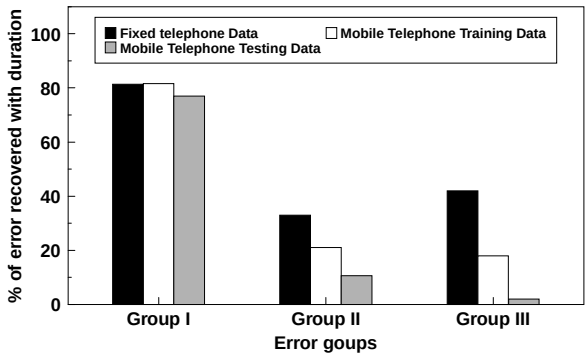

Figure 2: Duration Post-Processing Efficiency as Obtained on the Three Error Groups.

Figure 2 illustrates the pertinent phone-duration post-processing efficiency for the three remaining error groups. It appears that the highest error recovery is achieved for the first error group, the one, which contains the highest average error per number-pair. While for groups II and III there is a clear difference between results obtained on mobile data versus PSN data (results on mobile being worse) for the first error group no such difference is observed.

7. CONCLUSION

This pilot study showed the possibilities and the limits of duration post-processing in speech recognition when the phone duration is used to recover recognition errors. The most obvious limitation of this application is that HMM segmentation is often phonetically inconsistent. In fact, at the present stage of this study we can assume that if the segmentation has been more correct, the error recovery would be significantly higher.

A positive finding of the study is that the duration post-processing, though using often poorly detected phone duration, can help very efficiently in cases where spectral information alone is unable to propose the correct solution.

Other parameters can also be used for errors which can not be recovered using segment duration alone. One can envisage to use F0 movement to correct erroneous HMM segmentation when a confusion is made between numbers 0-800 and [zero-yitsə̃] and 08 [zero yit(ə)]. F0 movement can also be used in addition to phone duration in order to decide whether the syllable is stressed or not (error Group II).

8. REFERENCES

[1] Ostendorf, M. & K. Ross (1996), A Multi-level Model for Intonation Labels, Computing Prosody, pp. 291-308.

[2] Bartkova, K., D. Jouvet & T. Moudenc (1995), Using Segmental Duration Prediction for Rescoring the N-best Solution in Speech Recognition, *Proceedings of ICPhS95*, Stockholm, Vol. 4, pp. 248-251,.

[3] Chung, G. & S. Seneff (1997), Hierarchical Duration Modelling for Speech Recognition Using the Angie Framework. *Proceedings of EROSPEECH'97*, Patras, Vol. 3, pp.1475-1478.

[4] Bartkova, K. & D. Jouvet (1997), Usefulness of Phonetic Parameters in a Rejection Procedure of an HMM Based Speech Recognition System, *Proceedings of EURSPEECH'97*, Patras, Vol. 1, pp.267-269.

[5] Jouvet, D., K. Bartkova & J. Monné (1991), On the modelisation of allophones in an HMM based speech recognition system system, *Proceedings of EUROSPEECH'91*, Genova, pp. 923-926.

[6] Di Cristo, A. (1978), De la prosodie à l'intonosyntaxe, Thèse de Doctorat d'Etat, Université de Provence, Aix-Marseille I.