TWO-PASS UTTERANCE VERIFICATION ALGORITHM FOR LONG NATURAL NUMBERS RECOGNITION

Javier Caminero, Eduardo López (*), Luis Hernández (*)

Speech Technology Group, Telefónica Investigación y Desarrollo

Emilio Vargas 6, E-28043 Madrid, Spain

email: jcam@craso.tid.es

ABSTRACT

There are many Spontaneous Dialogue Recognition based applications like home-banking ones where long numbers recognition facilities are crucial to complete a request from the user. Rejection and Utterance Verification (UV) are difficult problems in these applications. In this contribution we improve our previously proposed UV procedure [3] in order to increase the correction of recognition errors, to solve grammatical ambiguities from the user, and to make more efficient the rejection of misrecognized or out-of-vocabulary (OOV) utterances. In spite of the verification performance, the proposed algorithm complies with the real-time constrains which are mandatory in real applications.

We evaluate our method and present recognition results from the long natural number recognition task of a real Data Driven application through the telephone line on a multilingual environment [2]. Experimental results show that the proposed method obtains a significant reduction in terms of recognition errors and achieves an extraordinary low false acceptance rate in all cases for different languages.

1. INTRODUCTION

Nowadays Spontaneous Dialogue Recognition Systems are demanding a growing interest. Human-machine interaction is becoming more and more natural and these systems must provide to the users with efficient long number recognition facilities.

Long number recognition allows complicated information transactions between a remote system and a user through the telephone line. Usually numbers with more than five digits are needed to provide Personal Identification Numbers (ID), credit card and bank account numbers or big money amounts, that will be referred to as long numbers. The recognition of these numbers is usually very important to achieve the successfully completion of an information interchange and these informations are difficult to be substituted with complementary data.

In this paper four major fenomena present when spontaneously pronouncing long natural numbers are considered: grammatical relaxation, word spotting, presence of noise and out of vocabulary utterances.

• **Grammatical relaxation**. In the three most spoken languages from Spain, which are the aims of this study, that is, Spanish, Catalan and Galician, and in many other languages, people usually pronounce numbers which are familiar to them like somebody's own ID number by splitting or grouping the figures in a way which makes the number easier to be pronounced or remembered. As, for example, *twelve thousand three hundred and ten* can be uttered as *twelve thousand three ten*. Although the uttered number is not strictly correct from a grammatical point of view, it can be understood by any other people. These grammatical relaxation must be corrected by the ASR system, that must be able to find out what number the user is referred to. It is important to allow the user to pronounce the number as he/she wants, so that the automatic system was accepted by the user as if it was an human operator.

• **Word Spotting**. Another spontaneous event produced when people pronounce long numbers is to surround the number with related application words like "my account number is ...", or the finish of the pronunciation of a money amount by naming the currency. In order to isolate the natural number recognition task from other tasks of the Dialogue Recognition System and due to the presence of this surrounding words we have to deal with them through the use of Word Spotting capabilities. Through these capabilities speaker hesitations must also be modelled. Such hesitations are very common because of the difficulties to pronounce and remeber very long numbers.

• **Presence of Noise**. In systems which operate through the telephone line, it is also very common the presence of several sources of noises, both channel noises and noises from the speaker himself or from his environment, like coughs, mouth noises, lips or breath noises, telephone clicks, etc. We have to consider them through what we have designated as Noise Spotting.

• **Out of Vocabulary Utterances**. Another important point to be considering in the design of real application is the rejection of OOV utterances, that is, those utterances which do not have any word which belong to the vocabulary.

In order to make a system robust against the four previous events we have developed and tested two complementary techniques: 1) word and noise spotting and 2) the use of an Utterance Verification (UV) procedure.

• **Word and Noise Spotting**. Looking for robustness against typical noises and task-dependent out-of-vocabulary words, we

(*) E.T.S.I. Telecomunicación. Universidad Politécnica de Madrid. Spain.

train specific Hidden Markov Models (HMM) to deal with them. Typical OOV words and noises, which appear surrounding the numbers, are obtained through the capture of specific data bases in real applications. In this way, through the use of robust recognition grammars, we provide our system with Word and Noise Spotting capabilities.

• **Utterance Verification** (UV). Although theoretically a single-step UV procedure [4] is optimal, we propose a robust two-pass UV algorithm suitable to integrate different information sources. The first pass is based on the evaluation of acoustic confidence measures from the Viterbi decoder, whose goal is to reject a complete pronunciation which does not belong to the vocabulary and comes from a non-cooperative speaker. The way to do that is by applying on-line garbage models which have been proposed in [1]. By this procedure we avoid the need of training specific HMMs to deal with a wide range of OOV, obtaining much better results. Note that the training of specific HMM models is only reserved to well-known task-dependent OOV words and noises for word and noise spotting. Furthermore the calculation of acoustic confidence measures in UV requires very few extra computational time and much less compared to the cost of the inclusion of specific garbage models in the recognition network. So that, this approach combines efficiency and reliability. As our system has to comply with the real time constrains, we use an endpoint detector to keep the recognizer inactive while the user pauses for a long time. By the use of the endpoint detector the utterance of a long number is split into separate pulses, so we have to obtain global or whole utterance acoustic confidence measures from partial or pulse ones. In the second step of the UV algorithm, confidence tests based on linguistic information are performed. The words to be recognized are separated into different linguistic categories and a rejection strategy is performed based on the coherence of such categories. This way of integrate different information sources in a recognizer is a widely used strategy in language understanding [5]. A correction of grammatical errors is also performed both from those which come from the spontaneous way a user pronounce a long number as from the deficiencies produced by the use of a too flexible recognition grammar in order to allow the partition of a pronunciation into an arbitrary number of separate pulses.

To evaluate the proposed methodology, we use a multilingual environment, with the three most widely spoken languages in Spain, that is, the Spanish, Catalan and Galician languages. To perform the evaluation we have used the VESTEL telephonic database speech corpus recorded by Telefónica I+D [7] for the Spanish language and another two telephonic databases which have been recently recorded to design Spontaneous Dialogue Recognition Systems in Catalan and Galician. These databases include a natural connected number corpus which has spontaneously spoken long numbers.

The rest of the paper is organized as follows: in Section 2, the baseline continuous natural numbers recognition system, including the test and training corpus, will be presented. In Section 3 the two-pass UV algorithm will be described in detail. Experimental results and conclusions will be given in Sections 4 and 5 respectively.

2. BASELINE SYSTEM

Our Continuous Natural Numbers Recognition task uses Continuous Hidden Markov Models (CHMM) with 18 Mel-cepstra parameters including the cepstra, delta- cepstra and the energy and delta-energy. The system is based on both word and sub-word models using a variable number of states per model. In the three languages to study, the recognition grammar presents a perplexity between 35 and 43, close to the vocabulary size, because most of the words can be followed by any other.

For the Spanish language, training and test sets have been obtained from the VESTEL database [7]. The corpus has been recorded by asking the caller to say its Identity Card Number, so that spontaneous answers from the callers could be obtained. A set of 4000 files was used to train the models and to perform recognition tests another different set of 1304 files was used. Both sets are balanced respect to the number of pronunciations of all dialectal zones of Spain and all the utterances belong to different speakers. The average number of digits of the test set is 7.8, but it has a variable number of words with arbitrary grouping.

For the Catalan language, training and test sets have been obtained from the VOCATEL telephonic database, a recently recorded database by Telefónica I+D and Catalonia Polytechnic University. The training set has 5000 utterances and the test set has 741 utterances with an average length of 6.92 digits. Utterances from both sets come from different speakers, and they were also obtained by asking the caller for his ID card number.

For the Galician language, the VOGATEL telephonic database has recently been captured by Telefónica I+D and Vigo University. We have employed 4500 utterances for training and 994 test utterances with an average length of 6.22 digits.

3. UTTERANCE VERIFICATION

As we have said before, our UV algorithm has two passes. In the first pass an acoustic confidence measure (ACM) is obtained in order to accept or reject the utterance. If the utterance has not been rejected a second pass consisting of linguistic processing is performed. Several linguistic categories are distinguished, these categories represent a set of units such digits, teens, hundreds, ..., that are needed to describe different spontaneous ways to pronounce long numbers. From the identification of the spontaneous structure of an utterance we rebuild a grammatically correct representation of it. Therefore grammatical flexibility from the user in his spontaneous way of pronouncing the numbers is interpreted through the introduction of external grammatical knowledge. As we will see, we can also correct some recognition errors produced by the use of our endpoint detector which stops the recognizer while the speaker does long pauses to optimize the system resources.

3.1. First Pass: Obtaining an Acoustic Confidence Measure

As we mentioned before, as our recognizer has been designed to work in a multichannel environment, an endpoint detector is used to split an utterance containing long pauses into several pulses between pauses. In this way no recognition resources are assigned to a channel during long silence gaps which are very frequent during the pronunciation of long numbers. For this reason our UV procedure starts with the evaluation of a partial Acoustic Confidence Measure (ACM) for each pulse k, P_ACM(k). This partial ACM is obtained from the average of the Viterbi probabilities of word and garbage models normalized by the pulse duration as it is described in [6]. It is important to note that each pulse can contain a sequence of N recognition units U_i, which can be vocabulary words or specific task-dependent OOV words or noises provided by the use of our robust word and noise spotting grammar. Each recognized unit inside a pulse will be extended over a frame interval t_i which has a duration of ΔU_i frames. Therefore the partial ACM for pulse k, P_ACM(k), can be obtained as:

$$P\_ACM(k) = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{\Delta U_i}\sum_{t\varepsilon t_i}\log[P(U_i/O_t)]\right\}$$

where $P(U_i/O_t)$ represents the Viterbi probability of unit U_i for the observation O_t from frame interval t.

In this way we reject those pulses which do not have an acceptable P_ACM, based on a rejection threshold, which are usually pulses with no vocabulary words and/or typical OOV words or noises. Ofcourse we also reject those pulses where the P_ACM is above the threshold but only contain OOV words or noises.

From the remaining unrejected pulses we obtain a global utterance ACM, U_ACM. The U_ACM is obtained by combining the N-best paths from the N-best recognized candidates and the L-best local scores from the frame-by-frame state probabilities of the whole set of HMM states. Probabilities for the N-best paths and the L-best local scores are combined through a weighting vector *w* obtained with Linear Discriminant Analysis (LDA), as we presented in [3].

$$U\_ACM = p^t w$$

where

$$p^t = \{P(O/W^1), \ldots P(O/W^N), P(O/S^1), \ldots P(O/S^L)\}$$

and

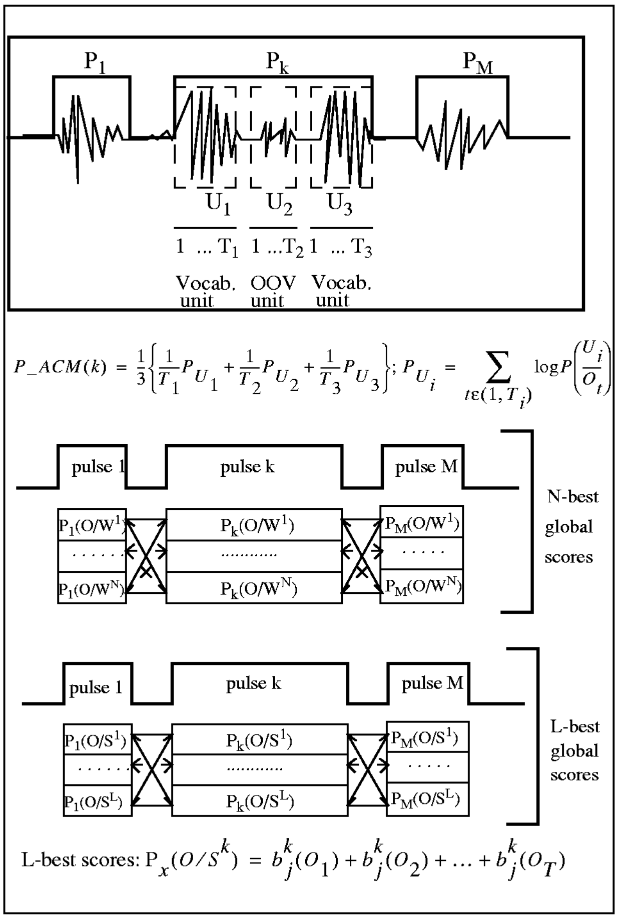
$P(O/W^k)$: ML scores of decoded hypotheses in the N-best list

$P(O/S^l)$: L-best scores

A decision threshold is applied to the obtained *U_ACM* value, and if the pronunciation overcomes this first acoustic verification pass, the second pass will be applied to it. Otherwise it will be rejected.

In the Figure 1 we can see a graphic explanation of this method.

3.2. Second Pass: Applying Linguistic Confidence



$$P\_ACM(k) = \frac{1}{3}\left\{\frac{1}{T_1}P_{U_1} + \frac{1}{T_2}P_{U_2} + \frac{1}{T_3}P_{U_3}\right\}; P_{U_i} = \sum_{t\varepsilon(1, T_i)}\log P\left(\frac{U_i}{O_t}\right)$$

L-best scores: $P_x(O/S^k) = b_j^k(O_1) + b_j^k(O_2) + \ldots + b_j^k(O_T)$

Figure 1: First pass of the algorithm: Obtaining the Acoustic Confidence Measure.

This second UV pass has several goals:

- The rejection of those utterances where no one from the N-best candidates can form an unique compact amount or number after applying the linguistic rules. We use the value of N to control the desired rejection level; the greater N, the lower rejection level, because it is easier to find a candidate which matches the rules in a longer candidate list.

- To apply those grammatical rules which could not be applied in the recognition network, because the recognition grammar is applied to each separate pulse and the number of pulses and the distribution of the keywords in the different pulses are arbitrary. Those utterances which do not overcome these complementary rules will be rejected.

- To apply spontaneous reconstruction rules, which allow the user to pronounce spontaneous ‘not-properly constructed’ long numbers, but which are understandable in communication between humans beings. In fact, in the VESTEL database, when we asked a person for his ID number, a 62% used ‘not-properly constructed’ long numbers. For example, for the long number 2,217,315 (two million two hundred and seventeen thousand three hundred and fifteen) if somebody says 2,000,000 217 315 (two million two hundred and seventeen three hundred and fifteen) a human listener can understand him although a not-properly constructed amount has been used. Of course not all groupings must be allowed, only those ones which are usually employed in each language. For the definition of those spontaneous reconstruction rules several categories are established: the hundred category (numbers between 0 and 999), the thousand category (numbers lower than a million formed by groups of the hundred category) and the million category (formed by groups of the thousand category).

• To reconstruct misrecognized utterances. In some languages some recognition errors are produced because the relaxation of certain ending sounds, which are crucial to distinguish e.g. between a full tens subgroup or two subgroups of a short tens plus a digit (e.g. 35 or 30-5). Another example is to distinguish between a full hundred group and a digit plus a single hundred subgroup (e.g. 325 or 3-125). All cases to be solved must have a different number of figures before and after applying the rules, therefore the external knowledge of the number of figures can be used to correct acoustic errors.

4. EXPERIMENTAL RESULTS

To show the performance obtained by the application of our two-pass UV algorithm, we perform recognition tests with the three most widely-spoken languages in Spain, that is, Castilian Spanish, Catalan and Galician. The long numbers construction rules for these three languages are quite similar, but there are some language-dependent peculiarities. The composition and size of the test sets have been clarified in the Section 2 of this paper (baseline system). We use a moderate rejection level (applied through the comparison threshold) in the first pass of acoustic confidence measures. In the linguistic second pass, we limit the number of candidates to be explored in the N-best list to N=3 (if we cannot find a candidate which matches all linguistic rules between the three first ones, the utterance will be rejected).

			Base	1st Pass only	2nd Pass only	1st+2nd Pass
Spanish	SER	1st c.	35.3%	31.0%	18.3%	15.7%
		2nd c.	25.4%	21.9%	12.6%	10.9%
	URR		0%	7.1%	6.4%	12.1%
	FAR		100%	21.3%	4.3%	1.5%
Catalan	SER	1st c.	17.3%	15.4%	13.5%	11.7%
		2nd c.	11.7%	9.7%	9.0%	7.3%
	URR		0%	4.1%	1.9%	5.5%
	FAR		100%	24.4%	4.6%	2.3%
Galician	SER	1st c.	18.5%	15.5%	13.4%	11.1%
		2nd c.	12.4%	10.1%	8.6%	6.8%
	URR		0%	5.1%	2.7%	7.2%
	FAR		100%	29.3%	3.9%	1.9%

Table 1: Results from applying the two-pass UV algorithm.

In Table 1 the results from applying pass by pass the two-pass UV algorithm are showed. Performance is evaluated through the Sentence Error Rate (SER) and the Utterance Rejection Rate (URR), that is, the rejection from correct utterances (in the base system the URR is 0% because no UV techniques were applied). SER for 2nd candidate were obtained by considering the two-best hypotheses which overcame the two-pass algorithm.

The lower performance obtained for the Spanish language is due to the use of an older database, which does not have training data enough for some words.

To evaluate the False Acceptance Rate (FAR) 878 OOV utterances, which were obtained from a real telephone application, have been used.

5. CONCLUSIONS

A new Utterance Verification procedure which consists of two passes has been presented in this paper. The combined use of acoustic measures and linguistic information has achieved an effective UV procedure with a low consumption of computational resources suitable to be applied to real telephone applications. Although linguistic information has been designed for the specific task of long natural number recognition, the simple way of applying and introducing it in the system makes the algorithm easily transferable to a wide range of applications. The critical and essential task of natural long numbers recognition in an Spontaneous Dialogue Recognition Application can be successfully solved by applying this algorithm.

6. ACKNOWLEDGEMENTS

We thank to the research groups of the Catalonian Polytechnic Univ. and Vigo Univ. for their assistance in the collection and elaboration of the databases for Catalan and Galician. The authors from ETSIT - UPM are also gratefull to CICYT for their partial finantial support under the contract no. TIC-96-0956-C04-03.

7. REFERENCES

[1] H. Bourlard, B. D'hoore, J.M. Boite, "Optimizing Recognition and Rejection Performance in Wordspotting Systems", Proc. ICASSP 94, pp. 373-376.

[2] J. Caminero, J. Álvarez, C. Crespo, D. Tapias, "Data-Driven Discourse Modeling for Semantic Interpretation", Proc. ICASSP 96, pp. 401-404.

[3] J. Caminero, L. Hernández, C. de la Torre, C. Martín, "Improving Utterance Verification Using Hierarchical Confidence Measures in Continuous Natural Numbers Recognition", Proc. ICASSP 97, pp. 891-894.

[4] S. Cox and R. Rose, "Confidence Measures for the Switchboard Database", Proc. ICASSP 96, pp. 511-514.

[5] S. Issar and W. Ward, "CMU's Robust Spoken Language Understanding System", Proc. EUROSPEECH 93, pp. 2147-2150.

[6] E. Lleida and R. Rose, "Likelihood Ratio Decoding and Confidence Measures for Continuous Speech Recognition", Proc. ICSLP 96, pp. 478-481.

[7] D. Tapias, A. Acero, J. Esteve, J.C. Torrecilla, "The VESTEL Speech Database", Proc. ICSLP 94, pp. 1811-1814.