# IMPROVING UTTERANCE VERIFICATION USING HIERARCHICAL CONFIDENCE MEASURES IN CONTINUOUS NATURAL NUMBERS RECOGNITION

*Javier Caminero, Luis Hernández (*), Celinda de la Torre, César Martín*

Speech Technology Group, Telefónica Investigación y Desarrollo
Emilio Vargas 6, E-28043 Madrid, Spain
email: {jcam, celinda, cma}@craso.tid.es, luish@gaps.ssr.upm.es

## ABSTRACT

Utterance Verification (UV) is a critical function of an Automatic Speech Recognition (ASR) System working on real applications where spontaneous speech, out-of-vocabulary (OOV) words and acoustic noises are present. In this paper we present a new UV procedure with two major features: a) Confidence tests are applied to decoded string hypotheses obtained from using word and garbage models that represent OOV words and noises. Thus the ASR system is designed to deal with what we refer to as Word Spotting and Noise Spotting capabilities. b) The UV procedure is based on three different confidence tests, two based on acoustic measures and one founded on linguistic information, applied in a hierarchical structure.

Experimental results from a real telephone application on a natural number recognition task show an 50% reduction in recognition errors with a moderate 12% rejection rate of correct utterances and a low 1.5% rate of false acceptance.

## 1. INTRODUCTION

Importance and popularity of Interactive Voice Response Systems are daily increasing. They allow complicated transactions and information exchanges between customers and remote information systems by simply using a telephone and their voices. In many cases, e.g. voice dialling or credit and account number based transactions, it is essential to provide the system with what can be called "long numbers recognition" facility: recognition of continuous natural numbers composed of more than five digits. In Spanish and in many other languages, see [1] for example, it is very extended to pronounce long numbers in many different ways, that is, to split a long amount into shorter groups of digits, tens, hundreds, thousands,..., to make it simpler to pronounce or easier to remember. Moreover it is also very usual to utter special words surrounding the number information, as for example "my number is...". And the acceptability of a service will be severely limited if the user is forced to change its speaking style.

All these factors make the long numbers recognition an important but also a challenging task in Continuous Speech Recognition. In this contribution, looking for an improvement in the performance of long numbers recognition, we will present a new utterance verification methodology developed to overcome three of the major difficulties we have found in real telephone applications:

- Rejection of Out-Of-Vocabulary (OOV) utterances: when all the words in a utterance are not included in the application lexicon (i.e. no long number is present).

- Long Number Spotting: recognition of long numbers embedded in typical surrounding words and hesitations (Word Spotting) or telephone noises (Noise Spotting).

- Spontaneous Speech: to deal with the wide range of spontaneous ways a customer can use to provide a long number.

The proposed methodology is based on the application of a hierarchy of confidence tests to solve the previous difficulties. The aim of our proposal is to combine acoustic confidence measures based on scores obtained from the Viterbi decoder, as proposed in [2], with confidence tests based on linguistic information. This will allow our system to be able to discriminate groups of words which are often confused with each other but belonging to different linguistic categories. The proposed confidence measures are applied following a three-step procedure that although theoretically suboptimal compared to a single-step procedure [3] is a robust strategy widely used in language understanding [4] to integrate different information sources in a speech recognizer. The major goals of the three confidence tests can be described as follows:

- The first test is based on an acoustic confidence measure at the world level. This measure is designed to deal with both Word and Noise Spotting, that is, to detect typical OOV words or telephone noises surrounding the long number information. Just because both typical OOV

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

words and telephone noises can be characterized with a reasonable degree of precision, a confidence measure is obtained by means of specific HMM models trained with automatically selected representative OOV words or noises.

- The second confidence test is also an acoustic one but now it is applied at the whole utterance level. In this case the goal is to reject a complete OOV utterance from a non-cooperative speaker. Therefore, as possible OOV words can belong to an open class, the confidence measure is based on the on-line garbage modelling proposed in [5]. Application of on-line garbage to utterance verification [6] avoids the need of training garbage models and thus the difficulty of training specific models for the high variety of acoustic possibilities in OOV utterances.

- Finally, a third confidence measure is obtained using linguistic information. In this case a rejection strategy has been developed based on the coherence of the reduced set of linguistic categories that can be used to construct a long number.

The evaluation of the proposed methodology has been made using the natural connected number corpus from the telephonic VESTEL database recorded by Telefónica I+D [7], which contains spontaneously spoken long numbers. A drastic 50% reduction in recognition errors is obtained with the proposed utterance verification procedure if compared to our baseline system without any utterance rejection facility. This improvement in recognition is obtained with only a 12% of Rejection Rate for correct utterances. The rejection capabilities of the system have also been improved achieving a 98.5% of OOV utterance Rejection at 1.5% of False Alarm.

In the following Sections of this paper we will provide some details on the baseline continuous natural numbers recognition system (Section 2) and on the evaluation and application of the three confidence tests we propose (Section 3). Experimental results and conclusions are also given in Sections 4 and 5.

## 2. BASELINE SYSTEM

Our Continuous Natural Numbers Recognition System use Semi-Continuous Hidden Markov Models (SCHMM) [8] with 18 Mel-cepstra parameters in three separated codebooks, for the cepstrums, delta cepstrums and the energy and delta energy. The system is based on both gender-dependent word and sub-word models using a variable number of states per model as it is described in [8]. The recognition grammar presents a perplexity of 43, close to the vocabulary size, because most of the words can be followed by each other.

Training and testing of the system is based on VESTEL database [7]. The corpus has been recorded by asking the caller to say its Identity Card Number, so far spontaneous utterances from the callers are obtained.

To train the models we use a set of 4000 files 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.

## 3. UTTERANCE VERIFICATION

An important point to understand the application of the proposed confidence measures is related with the way the system processes the incoming speech. Our system is designed to operate through the telephone line, and a real-time response is mandatory. Thus, in order to optimize the system resources and to increase the recognizer performance, we employ a pulse-based endpoint detector, which can filter big zones of silence or noise, allowing to de-activate the recognizer when no incoming voice is present. Therefore the pronunciation of a long number is usually split into several pulses. Then, as we will see, the confidence tests can be applied to both the pulse-level or the whole utterance-level.

### 3.1 Acoustic Confidence at Pulse Level

A first level in the hierarchy of utterance verification is applied at the pulse-level generally for groups of words. In this level we use explicit garbage models, since the garbage we try to model comes from two specific sources: the first one are typical telephonic noises, and the second source is the appearance of certain words, that use to accompany the pronunciation of a number. The rejection at this level allow us to include both Word and Noise Spotting. Pulse rejection is based on a single global threshold applied over the Pulse Acoustic Confidence Measure, *P_ACM*, obtained from the average of posterior probabilities of vocabulary and garbage models normalized by the pulse duration. These probabilities are provided by the speech decoder through the Viterbi algorithm. Equation (1) defines the *P_ACM(k)* for a pulse $k$ decoded as a sequence of N units $U_i$, words or garbages, $\{U_1, U_2, \ldots U_N\}$, each one of them extended over a frame interval $t_i$ with a total duration in number of frames $dU_i$:

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

Due to the degradation of the recognition accuracy in cases of a high presence of garbage models, we also impose a limit in the number of appearances from these

garbage models. If a maximum number is reached, we consider that the global quality of the pronunciation is not acceptable, and we reject it without any further information from the following confidence measures.

### 3.2 Acoustic Confidence at Utterance Level

In a second level, we perform a verification test for the whole utterance, using a method based on the information obtained 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, as we presented in [4]. Three recognition passes are necessary to obtain the verification score, a forward pass, where the acoustic probabilities of the N-best paths and the L-best local scores are computed and stored, and two backward passes, the first one to compute distance measures at the pulse level, and the second one to obtain the utterance verification score based on the information computed in the previous pass. The Utterance Acoustic Confidence Measure, *U_ACM*, is obtained as follows:

$$U\_ACM = p^t w \qquad (2)$$

where

$$p^t = \{P(O/W^1), \ldots P(O/W^N), P(O/S^1), \ldots P(O/S^L)\} \quad (3)$$

combines the ML scores of each decoded hypothesis $k$ in the N-best list $P(O/W^k)$ and L-best local scores obtained as:

$$P(O/S^l) = \sum_{s_t^l \in S^l} \log[P(O_t/s_t^l)] \qquad (4)$$

$S^l=\{s_1^l,\ldots, s_t^l,\ldots, s_T^l\}$ is the sequence of states which provides the best $l$ local scores.

The combination of the N-best global scores and the L-best local 1 scores is done through a weighting vector $w$ obtained with Linear Discriminant Analysis [4].

In this level, utterance rejection is based on a decision threshold applied to the obtained *U_ACM* value.

### 3.3 Linguistic Confidence

The third hierarchical level is based on linguistic information. According to our analysis over the training corpus of VESTEL database, when we asked a person for his Identity Card Number, a 62% did not used a single amount, but his own arbitrary digit groupings. In some cases, the amount was not correctly split, but it is understandable in communication between human beings. For instance, in the amount '1234456', some people could say '1-2-34-4-56' (one-two-thirty four-four-fifty six), which is only one of the many natural and correct ways of grouping, but other people could say any other "not-properly" constructed arbitrary digit groupings, for example '1000000-2-34-400-5-6' (one million-two-thirty four- four hundred-five-six), which can only be understand applying an error recovering strategy based on linguistic information.

For the linguistic confidence level, we distinguish several linguistic categories. A first one, the hundred-category, is related to the groups of numbers between 0 and 999. The second one is the thousand-category, which includes groups of the hundred-category to form numbers lower than a million. To form numbers bigger than a million (and lower than a million million) we establish a third category, the million-category, which include groups from the thousand-category.

The linguistic verification procedure works as follows:

- An N-best list is obtained from the recognition process.
- For each hypothesis in the N-best list the different linguistic categories are obtained.
- Depending on the categories that we have, we apply some reconstruction rules, which perform a verification procedure and an error recovering strategy. The aim of these rules is to obtain a "legal" and "compact" representation of the pronounced long number as a correct amount no matter the spontaneous groups of sub-amounts pronounced by the user.

Below is showed an example of how reconstruction rules can be applied to build an amount from an arbitrary set of groups either pronounced by the user in that way or as a result of some recognition errors which are typical in some dialects from Spanish, due to the relaxation or even disappearance of certain ending sounds which are crucial to distinguish between a full tens subgroup or a short tens subgroup plus a digit (e.g. 25 or 20-5). Of course, both error sources could be present in an utterance.

Pronounced ID n.: 32217345 (with relaxation of certain sounds)
Misrecognized ID n.: 30-2-1000000-217340-5
First reconstruction pass: 32000000-217345
Final reconstruction pass: 32217345

---

Pronounced ID n.: 32000000-217-3-45
Recognized ID n.: 32000000-217-3-45
First reconstruction pass: 320000-217345
Final reconstruction pass: 32217345

A utterance is rejected at this level if it was not possible to compact none from the N-best first candidates.

Therefore, depending on the number of candidates evaluated in the N-best list, the rejection level can be lower or higher.

### 3.4 Utterance Verification Procedure

The global rejection level of an utterance can be adjusted in each of the three passes. At the pulse level, a penalty can be applied to the garbage models to favour or not their appearance, and it can also be applied a maximum limit on

the number of gargage models allowed. In the acoustic utterance verification level, depending on the rejection threshold that we fix, we can reject or accept whole utterances. And at the linguistic level, the rejection level is controlled with the number of N-best candidates: the rejection level can be reduced by increasing N and increased by reducing the number of candidates.

These possibilities provide the system designer with a high flexibility to adapt the system to deal with very different possible situations. Experimental results for a particular task of Identity Card Numbers Recognition are presented in the next Section, but many other possible tasks could be easily handled.

## 4. EXPERIMENTAL RESULTS

In this section we present results for the proposed utterance verification procedure. The test corpus is composed of 1304 utterances, recorded from a real telephone application, with an average length of 7.8 digits (but a variable number of words). Rejection levels in each of the three passes were adjusted as follows: we set a low penalty factor at the pulse verification level to allow a high word-spotting capability and a low utterance rejection in this pass. A medium rejection level was used for the acoustic utterance confidence level. Finally, at the linguistic level, we limited the number of candidates in the N-best list to explore to N=3.

For the baseline system, without any utterance rejection capability, the Word Error Rate (WER) is 2.1%. In Table 1, we can see the Sentence Error Rate (SER) and the Utterance Rejection Rate (URR), which, obviously, is 0% in this case. In the Table, Sentence Error Rates for 2nd candidate were obtained considering as errors those cases where the correct hypothesis was not found in any of the two-best hypothesis that could survive the tree-level rejection system. Also in Table 1, we present the results obtained from applying each hierarchical verification level, and the results obtained after the application of the three levels.

| System | SER (1st cand.) | SER (2nd cand.) | URR |
|---|---|---|---|
| Baseline | 35.3% | 25.4% | 0% |
| Applying Rejection at Pulse Verification Level | 35.2% | 25.3% | 0.3% |
| Applying the Utterance Verification Score | 31% | 21.9% | 7.1% |
| Applying Linguistic Processing with N=3 | 18.3% | 12.6% | 6.4% |
| Applying the 3-previous verification passes | 15.7% | 10.9% | 12.1% |

Table 1: Results from applying hierarchical verification techniques

Evaluation of the rejection capabilities over OOV utterances was also tested using 878 OOV utterances, which were also from a real telephone application (VESTEL database [7]). A Rejection rate of 98.5% was obtained with a 1.5% of False Alarms.

## 5. CONCLUSIONS

In this paper we have presented a new Utterance Verification procedure based on the combination of acoustic and linguistic information. Acoustic information has been obtained through the use of garbage models for OOV speech and noise and from the best frame-by-frame local scores of the whole set of HMM sates. Linguistic information has been designed for the specific task of natural number recognition. Based on both acoustic and linguistic information three different confidence tests has been proposed. Although the combination on the three confidence tests could be done in many different ways, and this will be the subject of our future research, we have combined them into a sequential hierarchical structure. And from the experimental results, we can conclude that the combination of the hierarchical measures provides a robust system suitable for real telephone applications. Moreover, the proposed linguistic confidence measure can be applied to a wide range of applications.

## REFERENCES


[1] C.N. Jacobsen and J.G. Wilpon, "Automatic Recognition of Danish Natural Numbers for Telephone Applications", Proc. ICASSP 96, pp. 459-462.

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

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

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

[5] H. Boulard et al., "Optimizing Recognition and Rejection Performance in Wordspotting Systems", Proc. ICASSP 94, pp. 373-376.

[6] J. Caminero et al., "On-line Garbage Modeling with Discriminant Analysis for Utterance Verification", Proc. ICSLP 96, pp. 2111-2114.

[7] D. Tapias et al., "The VESTEL Telephone Speech Database", Proc. ICSLP 94, pp. 1811-1814.

[8] C. de la Torre et al., "Recognition of Spontaneously Spoken Connected Numbers in Spanish over the Telephone Line", Proc. EUROSPEECH 95, pp. 2123-2126.