# SETHOS: THE UPC SPEECH UNDERSTANDING SYSTEM

*Antonio Bonafonte, José B. Mariño and Albino Nogueiras*
*{antonio,canton,albino}@gps.tsc.upc.es*

Universitat Politècnica de Catalunya
c/Gran Capità s/n
08034 Barcelona (SPAIN)

## ABSTRACT

In EuroSpeech'95, we presented the first version of Sethos, the speech understanding system which has been developed at the UPC. In this paper some improvements are incorporated at different levels of Sethos: language model, models of the semantic units and acoustic models. These improvements increase the percentage of correctly decoded sentences from 60% to 80%. Some experiments are presented to evaluate the influence of each information source on the final performance. Furthermore, the computational cost is analyzed arriving to an important conclusion: the configuration which gives the best performance is also the less expensive. The reason is that as better is the modeling, narrower is the beam of the search.

## 1. INTRODUCTION

Speech understanding is one of the final goals of the speech recognition research in order to provide a friendly man/machine interface. Last decade, some works have appeared which try to solve this final problem for semantically constrained applications.

In the sequential scheme, speech is decoded in words by a recognizer and the output feeds an understanding system designed using natural language processing techniques. The main problem is that recognizers are not perfect and it is difficult to adapt the understanding system to deal with the corrupted data produced. Hence, the system usually sends the "not understood" message. One solution proposed is to use generalized recognizers which give not only the best sequence of words, but the *n-best* sequences. The problem with the *n-best* paradigm is that different hypotheses share usually the same meaning with high redundancy between hypotheses. Another point is that the recognizer performance could be improved if more information was present in recognition. If a sentence has not sense in an application it should not be considered during recognition. This is the reason why systems have appeared where the recognition and the understanding parts are integrated. However, the integration of two complex systems results in a much more complex one, leading to a search which could not be afforded. Consequently only a few of such systems have been proposed.

Recently in [1] and [2] a different approach has been proposed. It consists on defining an intermediate semantic language which is sequential with the input. The final interpretation should be done from a transcription in this semantic language. Being sequential with the input sentence, the constrains of this language can be easily integrated in the recognition system. Furthermore, if a *n-best* decoder had to be used, different hypotheses would have different meaning and a small *n* could be used.

Sethos has been developed for understand inquires to a Spanish geographic database defined in [3]. The results which are given refer to a subset of the database called *MINIGEO*. However, most of the information sources of Sethos are learned directly from examples making easy to adapt Sethos to other tasks, assuming that enough training data exists.

The paper is organized as follows: section 2 makes a brief review of the structure of Sethos. Section 3 explains the modeling improvements. Afterwards, section 4 evaluates the performance and complexity of different configurations of the system.

## 2. SETHOS: A BRIEF REVIEW

The aim of Sethos is to decode the speech onto an intermediate semantic representation. The semantic language is sequential with the input, that means that the first semantic unit is associated to the first words of the sentence, the second semantic unit to the following words, etc. Therefore, the semantic representation of each inquire is a string of semantic units which can be modeled by classic language modeling of symbol strings. In [4], a *trigram* was used to model the semantic language. The *trigram* probabilities were estimated from the semantic labels of 1000 inquires.

The second information source consists of models for each semantic unit (for instance *river*, *longer-than*, etc.). Each semantic unit appearing on the training database is associated to a sequence of words. In the baseline system, the words where transcribed phonetically and, for each semantic unit, an *n-gram* was inferred. The estimation was based on phones so that the grammar could learn the regularities of some relevant words which, due to inflections, share the same roots. Furthermore, in this way the system can accept any sentence without the definition of a lexicon. In practice, if some words appear which have not been seen during training then, as far as the words are not semantically relevant, the system will produce the correct semantic transcription. The phones of the unseen words will be parsed by transitions of the *n-gram* with low probabilities.

Finally, each phone is represented by a hidden Markov model. If the semantic-language model and the semantic units are represented as finite state automata, then all the automata can be integrated in a network. Thus, the problem of semantically

This work has been supported by CYCIT, contract TIC95-0884-C04-02

decoding speech can be viewed as the searching of the best path on the network given the speech observations. This problem can be solved using the Viterbi algorithm. The backtracking has to keep record of the semantic units which take part of each path. In fact, the implementation of the search algorithm does not use an integrated network because the size would be very large (and for more complex tasks it would become intractable). Instead, the different information sources are kept separately and expanded by demand when required.

# 3. IMPROVING MODELING

## 3.1. Semantic-language model

The first version of Sethos used *trigrams* to model the sequence of semantic units of the database inquires. The perplexity of the *trigram* over the test set was reduced using semantic classes. A semantic class was defined for each entity of the database: names of rivers, names of regions, numbers referred to length, etc.

*Trigrams* estimate the probability of a semantic unit given the two preceding semantic units. However, better performance and smaller complexity is obtained if *x-grams* are used [5]. *X-grams* compute the probability of a symbol based on the *x*-1 preceding symbols. The difference with *n-grams* is that the value of *x* is not fixed but depends on each particular situation. When *x-grams* are applied to model the semantic language, we obtain slightly smaller perplexity than with *n-grams* and also reduced complexities. Table 1 compares the perplexity for *bigram*, *trigram* and *x-gram*. The same table shows the complexity measured in number of states. This number comes from the representation of models as finite state automata. Each state is associated to a probability distribution function of the language model.

| ***Model*** | **perplexity** | **# states** |
|---|---|---|
| *bigram* | 7.63 | 94 |
| *trigram* | 5.81 | 780 |
| *x-gram* | 5.78 | 281 |

**Table 1**: Perplexity and number of states of the *FSA* needed to represent *x-grams* compared with *bigrams* and *trigrams*.

In this case the perplexity achieved by *x-gram* is the same than the achieved by *trigram.* However, only 36% of the parameters are needed. Most of the 281 states of the *x-gram* represent probabilities of semantic units given the previous unit or the two preceding units. However, some states represent longer histories: some probabilities are estimated using the five preceding semantic units. It should be noted that although the number of states is small, the reduction on the number of states is important because each state represents a semantic unit which can be represented by a complex model.

## 3.2. Models of the semantic units

As it has been discussed on the review, the baseline system expresses each semantic unit as a *n-gram* of phones. The number of available examples for each semantic unit is very small ranging from just one sample for some names, to 500 samples for the semantic unit *river*. The number of semantic units is around 90: 70 appear less than 25 times on the training data while only 7 appear more than 100 times. Fortunately, the more complex units are also those which appear more frequently on the data.

In [4] *bigrams* and *trigrams* offered almost the same performance. In this paper, *x-grams* will be used. In the case of semantic units with very sparse training data the resulting *x-gram* is very close to *bigrams*. However, in some cases, the memory of the model is much longer making possible to capture some frequent regularities, as for instance, the combination of phones which occurs in some frequent words.

The units that less appear on the training data are instances of attributes of the geographic database, for instance proper names of rivers and geographic regions. However, the way how these units are uttered is highly predictable. Furthermore, the different realizations of some semantic unit can be extended to other semantics units. For instance, some river names appear with prepositions on the training data while others do not. But the only reason is the sparseness of the training data. All river names can be said with or without preposition depending on the context. The second proposal for modeling the semantic units is to use graphs of phones for these concrete semantic units which can be selected from the relational scheme of the geographic database. A graph is built from the examples but the examples are generalized by hand. The human cost of the generalization is very low because the variability of these units is low.

Finally, the third technique which is applied to model semantic units is based on the definition of keywords (in fact roots) which are semantically relevant. Some complex units, as for instance *list*, can be expressed by a large number of words. However, some relevant words appear in most of the training samples. The idea is to define these words as one symbol. The *x-gram* is learnt from samples which are composed of phones and keywords. The same idea has been proposed recently for language modeling [8], but there, the training data was larger and the grouping of symbols was done automatically. Here, the lack of training data is supplied with *a priori* knowledge of the task.

## 3.3. Acoustic models

In the preliminary version of Sethos discrete HMM were used. Much better results are obtained if continuous HMM are used. As in [4], the acoustic data consist of mel-cepstrum coefficients, its first differences and the energy difference.

Furthermore, a recent study of Spanish sublexical units [7] shows how context dependent phones were as good as demisyllables and other syllabic units. Therefore, Sethos has been updated to use context dependent phones. However, because of the smoothing of probabilities, phones of the semantic models can be preceded and followed by any other phone. Therefore, it is not obvious the use of context dependent phones: the same situation occurs in continuous speech recognition when inter-word triphones are used. The current version of Sethos is only able to use context

dependent phones in keywords and in phone graphs. In these cases left context, right context or both are known independently of the recognition paths.

The acoustic models are trained from the Spanish Eurom.1 database [6] as was explained on [7]. Note that this database and the sentences which will be used to test the system are independent with respect the speakers, the text and the task. The selected context dependent phones are those triphones and rigth context dependent phones which appear more than 100 times on the acoustic training data.

# 4. EVALUATION OF SETHOS

In this section we propose some experiments to evaluate the performance of Sethos and also to establish the influence of each modeling technique which has been presented on previous section. Six configurations of the system are defined and evaluated on terms of performance as well as complexity.

## 4.1 Experiments definition

***Configurations* 1** and **2** compare the use of *x-grams* (1) with *bigrams* (2) to model the semantic language. Results with *trigrams* are not included because the performance is the same than *x-grams* but the complexity is greater. On these configurations semantic units are represented by *x-grams* and phones by context independent continuous HMM.

***Configurations* 3** and **4** analyze the benefit of using *a priori* knowledge on the design of semantic units. In particular, configuration 3 uses graphs of phones for instances of attributes. Besides that, configuration 4 uses the definition of semantic relevant roots. The semantic language is modeled by *x-grams* and the context independent continuos HMM are used to model phones.

***Configurations* 4** and **5** are similar to configurations 3 and 4 but context dependent phones are used when possible, as described above.

## 4.2 The performance results

The best way to evaluate the system would be accessing to a geographic database. Unfortunately, at the time of this paper we have no access to such database. Instead of that we dispose of 600 speech inquires along with the semantic transcription. In [4] it was noted that, in some cases, Sethos provides the correct transcription but different from the reference transcription; the reason is that the same semantic idea can be expressed in different ways on the semantic language. To have more accurate results, some inquires have been labeled with more than one transcription. Then, each decoded utterance is matched against the alternative transcriptions; the one which is more similar is used as appropriated reference and the performance is measured as the percentage between the correct semantic units over the length of the reference semantic transcriptions plus insertions. Furthermore, the number of inquires whose transcription is exactly the same than one of the annotated references is computed. Table 2 shows the results for the six configurations.

| ***Configuration*** | **% Correct (unit level)** | **% Correct (sentence level)** |
|---|---|---|
| 1 | 90.5 | 71.4 |
| 2 | 88.0 | 64.2 |
| 3 | 91.3 | 75.8 |
| 4 | 91.3 | 76.0 |
| 5 | 91.9 | 77.5 |
| 6 | 92.7 | 80.4 |

**Table 2**: Accuracy of the different configurations described on section 4.1.

The comparison between configuration 1 and 2 shows how *x-grams* are much better than *bigrams* as can be predicted by its lower perplexity. The error, either if it is measured in semantic units or in correct inquires, is reduced 20%.

Configurations 3 and 4 show how the use of a priori knowledge also improves the performance of the system. Configuration 3 uses phone graphs for some instances: it can be observed how the error is reduced in 15% (at the sentence level) with respect of using always *x-grams*. The reasons are two: first, as the structure of the graphs is more rigid than *x-grams*, the insertions and substitutions are reduced. On the other hand, the use of generalized examples, improves the quality of the models.

On the other hand, the use of keywords (configuration 4) has no effect on the performance. The reason must be that *x-grams* are able to capture the regularities which we pretended to capture with this technique.

Finally, configurations 5 and 6 show the results when context dependent units are used. The good properties of these units make the use of keywords becomes relevant, reducing in 31% the error at the sentence level. It should be remarked that the actual version of Sethos only makes use of context dependent units on keywords and graphs. The good results obtained makes desirable to extend Sethos so that context depending phones can be used in all the situations.

The result obtained with the best configuration (conf. 6) (92.7% of correct semantic units, 80.4% of correct sentences) compares very favorably to the results obtained on the first version of Sethos (86.3% and 60.0%). The differences are due to better modeling but also to the use of several reference transcriptions. In order to give a fair comparison it should be said that the results of configuration 6 when only one reference transcription is used is 90.7% of correct semantic units and 71.6% of correct sentences.

## 4.3. Evaluation of computational cost

Last section has shown how improving the modeling produces an improving of the performance: *x-grams* are better than *bigrams*, context dependent units are better than the context independents

ones, etc. The aim of this section is to evaluate the prize to be paid for getting such improvement.

The algorithm which has been used to decode semantically the inquires is the Viterbi algorithm. The models are not integrated in a unique network but maintained separately on memory so that memory needed to represent the models is low. In order to reduce the complexity, for each speech frame the search is limited to a beam near to the best hypothesis of that frame. Furthermore, for each active hypothesis at time *t*, new hypotheses are generated dynamically for *t*+1. Therefore, the beam width limits not only the computational cost, but also the memory required by the algorithm.

The beam of all the experiments presented on the preceding section has been defined by the same threshold. Table 3 shows the total number of hypotheses which should be explored if the search was not limited to a beam. In the same table, the mean number of active hypotheses for each speech frame is presented.

| ***Configuration*** | **# total hypotheses** | **# active hypotheses** |
|---|---|---|
| 1 | 18K8 | 219.0 |
| 2 | 5K4 | 224.1 |
| 3 | 21K2 | 238.0 |
| 4 | 24K7 | 239.3 |
| 5 | 21K2 | 191.8 |
| 6 | 24K7 | 186.3 |

**Table 3**: # of hypotheses per frame which should be explored without applying beam search compared with the # of hypotheses which have been explored to get the results of table 2.

First, it has to be noted that the number of total hypotheses, in the case that the semantic language is represented by a *x-gram* (conf. 1) is more than three times the number in the case that a *bigram* is used (conf. 2). The number is still larger if semantic units are represented by units with more structure as phone graphs (conf. 3 and 5) or phone graphs and keywords (conf. 4 and 6). However, note that the number of transitions between states is significantly smaller because some states of the automata can be acceded only from one state (for instance, if keyword *río* (river) is defined, the intermediate phones of */R//i//o/* can be acceded only from the previous phone of the word.

Another point to be noted is how the number of hypotheses explored is very small compared with the complete size of the network. Only around 1% of the search space is explored. Some experiments increasing the explored space produce similar results showing that the size of the beam was appropriated.

Finally, the most important conclusion is that the improvement of the modeling has not influence on the computational cost. For instance, note how the number of hypothesis explored with *x-grams* is even smaller than those explored with *bigrams*. The reason is that if the models are improved, then the best hypothesis becomes even better and therefore the beam becomes narrower. The same effect can be observed when context dependent phones are used. In all cases, as the performance of the system increases the beam becomes narrower. This is an important observation which can be applied to other improvements on speech modeling. Although configuration 6 needs to compute more gaussians probabilities, the time needed to perform the search is smaller than in the rest of configurations.

## 5. SUMMARY

In this paper we have presented Sethos, the speech understanding system which has been developed at the UPC: Sethos uses *x-grams* for modeling the semantic language and each semantic unit. However, better results are obtained if *a priori* knowledge of the task is incorporated in models of the semantic units.

Although the current version of Sethos only allows the use of context dependent phones in some cases, the use of these models improves significantly the performance. This motivates to extend Sethos so that context dependent units can be used in the rest of cases. Sethos is able to decode semantically 92.7% of the semantic units, producing 80% of the inquires perfectly decoded.

An important conclusion is that, because of the use of beam search, if better modeling results in an improvement of the performance, then the amount of time and memory required to decode the inquire decreases.

## 7. REFERENCES


1. Natividad Prieto et al, "Continuous Speech Understanding based on Automatic Learning of Acoustic and Semantic Models", *Proc. of ICSLP-94*, pp 2175-2178, Yokohama 1994

2. R.Pieraccini and E.Levin, "Stochastic Representation of Semantic Structure for Speech Understanding", *Proc. of EuroSpeech-91*, pp. 383-386 . Genova 1991.

3. F. Casacuberta, et al, "Development of Spanish Corpora for Speech Recognition for Speech Research", *Proc. of Workshop on International Cooperation ans Standarization of Speech Databases and Speech I/O Assesment Methods*. Chiavari. Septembre 1991.

4. A. Bonafonte et al, "Semantic decoding of Speech in Constrained Domains", *Proc. of EUROSPEECH-95*, pp. 559-562, Madrid 1995

5. A. Bonafonte et al, "Language Modeling Using *x-grams*", *Proc. of ICSLP-96*, Philadelphia, October 1996.

6. A.Moreno, *EUROM.1 Spanish Database*, Esprit Technology Assesment in Multilingual Applications. Esprit Project 6919. Report D6

7. A. Bonafonte et al, "Study of Subword Units for Spanish Speech Recognition", *Proc. of EUROSPEECH-95*, pp. 1607-1610, Madrid 1995

8. H. Masataki and Y. Sgisaka, "Variable-Order n-gram Generation by Word-Class Splitting and Consecutive Word Grouping", Proc. of ICASSP'96, Atlanta, 1996.