SYNTAX-BASED SPEECH RECOGNITION:
HOW A SYNTACTIC PARSER CAN HELP A RECOGNITION SYSTEM

Arnaud Gaudinat, Jean-Philippe Goldman, Eric Wehrli
LATL - University of Geneva
http://latl.unige.ch
{gaudinat,goldman,wehrli}@latl.unige.ch

ABSTRACT
This contribution is a report on an on-going research aiming at the development of a speech recognition system for French, combining a standard HMM recognition tool with a syntactic parser. Because of the very high number of homophones in French and because several agreement rules spread over an unbounded number of words, we designed our GB-based morphological and syntactic parser to output a correct orthographic form from a lattice of phonemes produced by the front-head HMM recognition system.

This resulting lattice of phonemes is processed by the syntactic parser, which selects the best word sequence according to its linguistic knowledge. The originality of this approach is the use of a syntactic parser, tuned to phonetic inputs, for a speech recognition task.

INTRODUCTION
Parsing a phonetic string into orthographic words is still a major problem in speech recognition. (Grosjean and Frauenfelder, 1996, McQueen et al, 1995). The main difficulty is the lack of information on the word boundaries. Two other morpho-linguistic reasons preventing the production of a correct orthographic sentence could be: the number of homophone words in French (recall that several of the most common grammatical affixes have no phonetic realisation in French) and the morpho-phonologic adjustments within words and at word boundaries, that affect the phonetic sequence (French schwa-elision and liaison for instance) (Dell 1983). These three problems combined with the low accuracy of automatic acoustico-phonetic decoder explain the poor results in speech recognition using such a method.

Current statistical methods based on HMM considerably improved the recognition rates. But major errors remain and final results are still not satisfactory. No explicit linguistic knowledge is used in such systems, as the best word sequence is found with statistical language modelling based on word N-grams. However, some phonologic and linguistic phenomena like agreement rules can spread over many words. Such distance is too wide for statistical methods and error in the final orthographic sentence can hardly be avoided.

To illustrate the agreement rule on number, consider the following example:

le chat et le chien semblent être contents

This is how should be interpreted the phonetic sequence:

le chat et le chien sembles être contents

in which grammatical errors occur on the verb semblent (seem) and/or the adjective content (happy). Addition of a relative clause modifying the subject, rapidly increases the number of words between the subject chat et...
chien and the words that must agree with them, semblent and contents (i.e. Le chat et le chien que j’ai adoptés durant le mois dernier semblent être contents - The cat and the dog that I have adopted last month seem to be happy)

In this example, 155 words can be extracted from the 21-phones unsegmented sequence according to our 200’000 words phonosyntactic lexicon. Such a number of candidates clearly shows the high level of ambiguity within a non-segmented phonetic string.

We describe here how the syntactic parser(i) segments the phonetic string to convert it into a word chart, using a lexicon of about 200’000 entries and a lexical analyser, (ii) filters out ungrammatical combinations of words and phrases with the syntactic parser, which builds all the possible constituents on the basic of the words, and (iii) is able to handle phonological adjustments at word boundaries, such as liaison, elision and denasalisation.

Finally we evaluate the robustness of the parser, along with its ability to recover from slightly altered phonetic inputs (deletion, insertion, substitution of phonemes). This was done by simulating various types of errors and various rates of errors within a 50-sentence phonetic corpus. Such an evaluation provides important information to determine the optimal width of the lattice computed by the recognition module.

**THE SYSTEM**

The system consists of two modules: the **lexical analyser**, which reads the phone sequence and produces a lattice of lexical items, and the **syntactic parser**, which builds grammatical syntactic structures on the basis of the word hypothesis.

The phonetic lexicon is built as a phonetic tree, where nodes are possible phones in a word and terminal nodes represent the final phone of a word and contain all the grammatical data of the lexical item.

While the incoming phone sequence is read from left to right, each phone is used to build a possible word according to the phonetic tree. Once a candidate is completely recognised, it is sent to the syntactic parser.

Consider the following phone sequence **leptipa** which can be parsed as in next figure.

We can see how many possible words can be candidates of the final sentence.

The **syntactic parser** (Laenzlinger Wehrli 1991 1993) tries to associate incoming structure to existing structures, like S-structures of the Chomskian Government-and-Biding Theory. It is made up of distinct modules. Some of them build structures (X’-module, chain-module, agreement module), some others work as filters and reduce the number of items (morphological/abstract case, thematic interpretation). All the modules are called after every word entry into the chart. Hypotheses are built in parallel and some heuristic rules filters out the flow of hypothesis to keep the best ones.

Elision and liaison phenomena are also considered. Syntactic constraints determine the pronunciation of a schwa or a latent consonant. If the elision or the liaison is optional, both cases are considered. For example, in the **leptipa** example, the lexical item petit (**pɔt̥i**) was recognised because elision rules agreed and phonetic realisation as **pti** was possible.

Phonetic sequence can be totally ambiguous as in the following example :

**ilretournasamezɔ**

*il retourna sa maison*

(he turned his house upside down)

**il retourne à sa maison**

(he gets back to his house)

**ildonynpwaŋedəmē**

*il donne une poignée de mains*

(he shakes hand)
Il donne une poignée demain

(he gives a handful tomorrow)

In this case, an heuristic module will allocate a better score to structure according to the attachments and movements.

In order to have an adaptable phonetic entry and to be able to cope with phoneme recognition errors, variants of pronunciation can be included in the input phonetic string. On the basis of this kind of phoneme lattice, the lexical analyser build all the possible lexical items words and feed the syntactic parser with them.

**EXPERIMENT**

The aim of this experiment is an evaluation of the behaviour of the parser with phonetic input, based on a simulation of phonetic data produced by a non-perfect phoneme recogniser, i.e. with phoneme mismatch, insertions and deletions.

We created a small corpus of 50 simple sentences (from 3 to 14 words). After an automatic grapheme-to-phoneme conversion, a manual verification was done in order to have a correct phonetic transcription of the 50 original sentences.

As we wanted to simulate separately various type of errors in phone recognition (deletion, insertion and substitution), we derived four series of phonetic corpora from the original one: three series with one of these error types and a fourth one in which the three types of errors are combined. Within each serie, we regularly increased the ratio of phonetic errors from 1 to 50%. So, for the corpus with the highest noise, one phoneme out of two has an error while in the 30%-corpus, about one phoneme out of three has an error.

Besides that, for the substitution corpus, we also varied the weight of the mismatch in terms of different phonetic variants added to the lattice. We added from 1 to 7 closest phonemes in term of distinctive features.

For example, here is a sentence, with its original phonetic transcription and an example of phone lattice.

*Le chat et le chien semblent être contents*

\[\text{lɒfælɒʃjësiblietkɒtə}\]


deletion substitution-4 insertion

We used the system to recognise the corpus with every type of error and at every phone error rate.

**RESULTS**

Figure 1 shows the behaviour of the system for substitution errors from 1 to 50% , and with 1 to 7 errors). We can see that (i) this type of error has dramatic effect on recognition if the rate of phone mismatch increases and (ii) a higher number of variants within a phoneme mismatch does not have a major effect on the recognition rate (see the superposition of the 7 lines).

Figure 2 shows general results for insertion, deletion, substitution with 5 variants and a combination of these three type of errors. We note that the system is able to handle with insertions of phonemes, but has poor results with deleted and substituted phones. The heavy line shows the system results with all types of errors.

An explanation of such differences between error types could be the correlation of the phonetic sequence and the number of lexical candidates. Small calculations showed that alternative phoneme sequences with insertion or deletion errors generate a lower number of lexical candidate than substitution errors. We can add all these candidates work as
competitors and create many syntactic structures that are filtered out. Thus, at this point, we could build a full speech recognition system composed of a phoneme recogniser and our system. On the assumption that (i) phone recognition error rate is about 20% (ii) the phone recogniser produces a lattice of phone with the 5-best solutions and (iii) the correct phone is in this set of 5 phones (i.e. the good solution is in the lattice) we should expect a word recognition rate of 70 %.

CONCLUSION

We present here a new method that uses explicit linguistic knowledge for a speech recognition task by validating hypothetical phonetic sequence with a syntactic parser. This parser together with a phonetic database reads the phonetic entry phone by phone, extracts all the possible lexical candidates and filters out syntactic structures as soon as they are non-valid. The system should always find a good solution on the assumption that the correct phonetic sequence belongs to the phonetic lattice. Morpho-linguistic phenomena like elision and liaison are fully treated. These rough results can still be improved with a greater robustness of the parser.

The next step should be the integration of statistical data such as word frequencies or statistical language modelling. The combination of it together with the linguistic approach proposed in the paper, could be a promising solution for speech recognition.

REFERENCES


The authors would like to thank the Fonds National de la Recherche Scientifique (subsidie n°12.50944.97), the Département de l’Instruction Public and the University of Geneva.