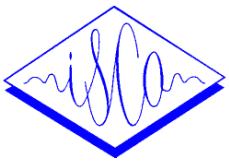

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

KNOWLEDGE-BASED PHONEME RECOGNITION

D. Ederveen, L. Boves

Nijmegen University, dept. of Language and Speech / PTT Research NT, TWS C47
Postbus 421, NL-2260 AK Leidschendam, tel: +31 70 3323202, fax: +31 70 3326477
e-mail: ederveen@lett.kun.nl / boves@lett.kun.nl

ABSTRACT / RESUMO

English: We are building a knowledge-based phoneme recognition system. After a short introduction to the system, and a discussion of some theoretical motivations, we present some results from the first stage: the recognition of the voiceless plosive class and the voiceless fricative class.
Keywords: speech recognition, acoustic phonetic knowledge.

Esperanto: Ni konstruas sciobazitan fonemrekonilon. Post mallonga enkonduko al la sistemo, kaj pritrakto de kelkaj teoriaj motivoj, ni prezentos kelkajn rezultojn de la unua stadio: la rekono de la senvoĉa ploziva kaj la senvoĉa frikativa klasoj.
Titolvortoj: parolrekono, akustikfonetika scio.

10.21437/Eurospeech.1991-107

INTRODUCTION

To knowledge-based speech recognition, there have been a number of approaches [1] [2] [6]. We want to acquire acoustic phonetic knowledge of the Dutch language. To achieve this goal, we are building a preprocessor that generates a phoneme lattice from the acoustical speech signal. The knowledge is represented as Prolog-rules. This phoneme lattice can then be used as the input for the lexical module of a speech recognizer. The acoustical parameters in a database with labelled Dutch speech *(DEMSI)*, which has been developed by the speech group of PTT Research [7], are the basis for this research project.

SYSTEM OUTLINE

The internal structure of the database is depicted in Figure 1. DEMSI contains for each speaker 101 single-syllable meaningful words, 101 single-syllable meaningless words, 545 meaningless words with two syllables, 28 meaningful sentences and one short text containing seven sentences. These utterances have been selected to constitute a phonetically balanced set. At present, we have transcribed and hand labelled the utterances of two male professional speakers, and added them to the database. The utterances of two additional speakers have been transcribed. This material will be labelled

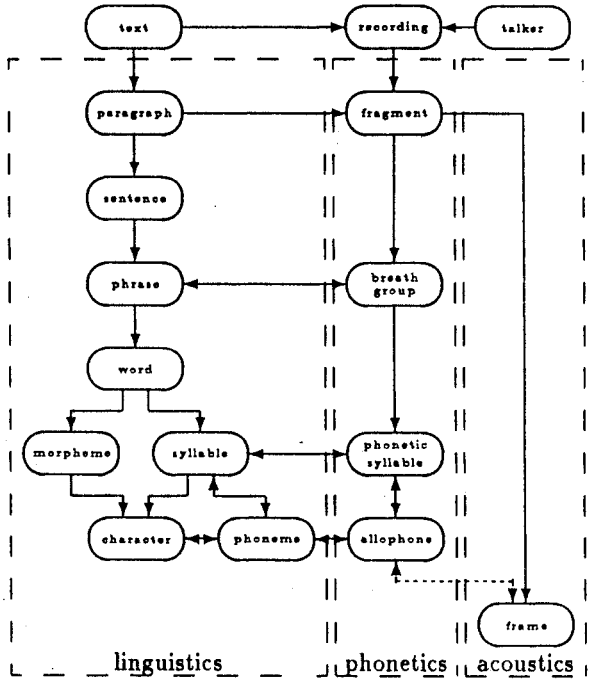

Figure 1: Internal structure of DEMSI

and added within a few months, and will be used for the final test of the recognizer.

This database, which is the primary source of knowledge, is the basis of the project. It contains e.g. labelling information and acoustical parameters, and can be searched for information about the behaviour of those parameters in different contexts.

The phoneme recognition system itself is divided into two stages. First, rules are used to make a segmentation based on manner of articulation. The output of this stage is a lattice of *broad phoneme classes* with a maximum depth of three. Associated with this lattice are measures of certainty, indicating for every recognized class the confidence with which that class can be recognized in that place. These certainty measures are derived from the presence or absence of (combinations

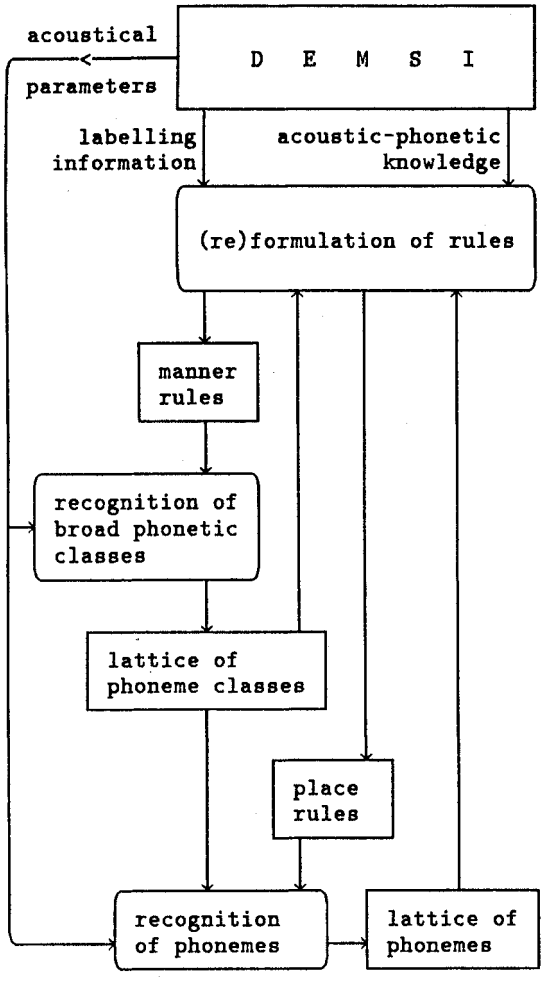

Figure 2: Development of the recognition system

of) acoustic cues.

In the second stage, rules are used to make a finer distinction, based on the information from the first stage, and on place of articulation. The output of this stage is a lattice of *phonemes* with a maximum depth of ten. In this lattice, the phonemes also have associated measures of certainty. Figure 2 displays the development of the system. The final system operates on the speech signal instead of on DEMSI, and operates without the box `(re)formulation of rules`.

MOTIVATIONS
for recognition in two stages

The division of the recognition system in recognition of broad phonetic classes (by means of manner rules) and recognition of phonemes (with place rules) raises the question why this has been done. There are a number of reasons for this.

Firstly, there are some practical reasons. Breaking a big problem into smaller pieces makes it less discouraging and more tractable. In addition, it is known that it is possible, at least for longer words, to perform word recognition on the basis of broad phonetic classes [4]. This has also been studied for the Dutch language [8], [9]. With this method it is possible, also for rather large lexicons, to reduce the search space considerably by means of a limited number of phoneme classes. After this reduction, the few remaining possibilities are checked and disambiguated relativily easily with cues for place of articulation. In this way, the results of both the first and the second stage can be used by a lexical or syllabic module. An important condition, however, is a reliable recognition of phoneme classes.

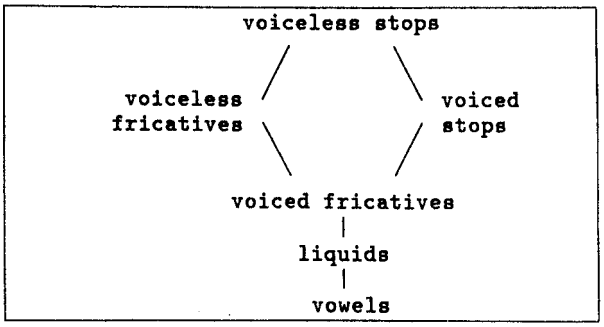

Figure 4: Lenition hierarchy

Secondly, there is a theoretical motivation. Within the model of *autosegmental phonology*, the set of phonological features can be represented in terms of a *feature geometry* [3], as shown in Figure 3. This tree structure clearly shows a distinction between features of the so-called *categorial gesture* (under the categorial node) and those of the *articulatory gesture* (under the articulatory node). This distinction corresponds to the division made in the recognizer. Although there is a manner node below the categorial node, and a consonantal (or place) node below the articulatory node, we will show below that the features, belonging to the categorial and articulatory nodes correspond to the features that can be recognized by the distinction between manner and place of articulation in terms of the phoneme recognizer, respectively.

for recognized classes

In the remainder of this paper, the focus will be on the first stage, the recognition of phoneme classes. In this first stage, we distinguish the following five categories: explosion, frication, voice, nasality and liquids/vowels. Based on these categories, we formulate rules for the recognition of the following broad phonetic classes: voiceless stops, voiceless fricatives, voiced stops and fricatives, nasals, and finally the group of liquids and vowels. The rules are being developed in the above order.

This order is remarkably consistent with the historical weakening or *lenition* of stops to vowels. On the

RESULTS

The performance of the rules for LP and LF is listed in Table 2. They have been tested with the 545 double-syllable words of two speakers. It should be noted that these are recognition results of two classes. To obtain a complete segmentation, we combine the recognition results of all six classes with their associated measures of certainty to one lattice of broad phoneme classes with a maximum depth of three.

Segments that are correctly recognized as belonging to the class, as well as segments that are correctly recognized as not belonging to the class, are calculated in the *performance*: $\frac{100c}{t}(c = correct, t = total)$. The *accuracy* also takes false alarms (f) into account: $\frac{100(c-f)}{t}$.

In our system, it is of great importance that *perf* is high, because missed segments cannot be recovered. On the other hand, *acc* may be lower, because false alarms can be weeded out by means of the measures of certainty, when combining the results to a lattice. The measures of certainty of correctly recognized segments are an average of 20 percent higher than those of the false alarms. Although *perf* doesn't seem very high, we note that, for example in the case of LP, the eight percent of missed segments also includes many word initial segments and unrealized plosives.

Because the scoring algorithm was designed to accomodate more than one class, it isn't very informative in the one-class case. If, for example, the sequence /sls/ is recognized as one LF segment, the scoring algorithm counts one correct segment /s/, but also two missed segments: the /l/ and the following /s/. This could also be counted as one correct and no missed segments, bearing in mind that the rules for liquids and vowels will find the /l/. Likewise, if the sequence /tp/ is recognized as two LP segments, the scoring algorithm counts one correct segment /tp/, but also two false alarms: the gap between the two recognized LP segments, and the second LP segment. This could also be counted as one correct segment and no false alarms, bearing in mind that it will not make a difference for a search in a lexicon with phonetic classes.

To show the difference, two figures are given: the first number in the column is the figure from the scoring algorithm, the second number in the column is the one-class figure. The last part of the table presents the results for LF when /x/ and /χ/ are not considered.

CONCLUSIONS

We have tried to show that for the distinction between manner and place of articulation there is a phonological justification in the form of feature geometry. Similarly, the proposed order of recognition corresponds to an order that ranges from more to less salient, thus ensuring that the most reliable recognition is done first.

Besides, we hope to have made clear that there are still other promising possibilities besides neural networks and HMM. An explicit representation of acoustic phonetic knowledge helps understanding the process of speech recognition, makes it possible to perform knowledge-based phoneme recognition, and can eventually be used in realtime systems.

	speaker 1		*speaker 2*		*both speakers*	
LP						
perf	91.7	91.7	91.8	91.8	91.8	91.8
acc	51.1	53.4	38.5	41.1	44.8	47.3
LF						
perf	83.3	85.1	82.0	82.2	82.6	83.7
acc	76.7	78.5	65.8	66.1	71.2	72.3
LF†						
perf	93.0	95.2	92.2	92.5	92.6	93.8
acc	80.7	82.9	70.3	70.5	75.5	76.6

† without /x/ and /χ/

Table 2: Performance and accuracy for voiceless plosives (LP) and voiceless fricatives (LF), with the general and the one-class scoring algorithm

References

[1] ALTOSAAR, T., AND KARJALAINEN, M. A knowledge-based approach to unlimited vocabulary speech recognition for the finnish language. In *Proceedings Eurospeech* (Paris, France, september 1989), vol. 2, pp. 613–616.

[2] BULOT, R., AND MÉLONI, H. Processing acoustic and phonetic knowledge in prolog. In *Proceedings of the 3rd International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA)* (Varna, Bulgaria, 21-23 sept 1988), T. O'Shea and V. Sgurev, Eds., Elsevier Science Publishers BV (North-Holland), pp. 177–185.

[3] CLEMENTS. The geometry of phonological features. In *Phonology Yearbook 2*, C. Ewen and J. Anderson, Eds. printed in Great Britain, 1985, pp. 225–252.

[4] HUTTENLOCHER, D. P., AND ZUE, V. W. A model of lexical access based on partial phonetic information. In *Proceedings ICASSP* (1984), p. 26.4.

[5] LASS, AND ANDERSON. *Old English Phonology*. Cambridge University Press, 1975.

[6] MIZOGUCHI, R., TSUJINO, K., AND KAKUSHO, O. A continuous speech recognition system based on knowledge engineering techniques. In *Proceedings ICASSP* (1986), pp. 1221–1224.

[7] VAN HEUGTEN, B., HENDRIKS, J., BOVES, L., VAN ERP, A., HOUBEN, C., AND VAN GOLSTEIN BROUWERS, W. Proceedings 1987 of the speech research group. memorandum 1631 DNL/88, PTT Research Neher Laboratories, Leidschendam, The Netherlands, 1988.

[8] VEENHOF, T., AND BLOOTHOOFT, G. Statistics of sequences of broad phonetic classes in newspaper dutch. Progress Report 1, Institute of Phonetics, Utrecht, The Netherlands, 1987.

[9] VERNOOIJ, G., BLOOTHOOFT, G., AND VAN HOLSTEIJN, Y. Simulation of isolated word recognition on the basis of a hierarchy of phonetic classes. In *Proceedings Eurospeech* (Paris, France, september 1989), vol. 2, pp. 469–472.

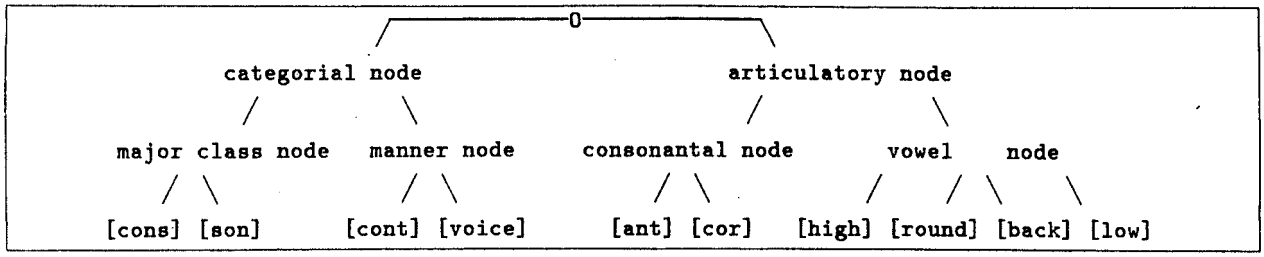

Figure 3: Feature geometry

basis of this process a *lenition hierarchy* [5] has been established, as given in Figure 4. From top to bottom it shows the weakening process. This is an indication that the proposed order of formulation of rules is indeed an order which begins with the easy part and leaves the frills for desert. In other words, those phoneme classes that are most salient, and therefore recognizable with the highest measures of certainty, are first dealt with. In that way the last and most difficult phoneme class (vowels and liquids) can be identified with a relativily high measure of certainty, as they are simply the remaining, unidentified parts of the speech signal.

With the six classes mentioned above, it is possible to distinguish between the four features of the categorial node in Figure 3. For example, in Table 1, the feature *[voice]* can be distinguished by separating the voiceless stops and fricatives from the rest of the classes found. Thus, these six classes can determine all features of the categorial node of Figure 3, and shows that the recognition of these classes is a complete recognition in phonological terms.

This contrasts the (not uncommon) approach of building feature detectors for the features *[son]*, *[cons]*, *[cont]* and *[voice]*, and afterwards combining them to recognize classes of phonemes.

RECOGNITION OF CLASSES

The recognition of a particular class in an utterance from the database is done as described below. At first, the database is opened and the utterance is selected. Then, Prolog backtracks through the rules that apply to that class, and combines the lattices delivered by the rules to one. In every rule, the acoustical parameters that are used are selected, and if necessary loaded from DEMSI and asserted into the Prolog database as a series of facts, for example:

	[son]	*[cons]*	*[cont]*	*[voice]*
voiceless stops	−	+	−	−
voiced stops	−	+	−	+
voiceless fric.	−	+	+	−
voiced fricatives	−	+	+	+
nasals	+	+	−	+
liquids/vowels	+	−	+	+

Table 1: Features and phoneme classes

```
param(Utterance, Frame_Nr, Value).
```

If the rule requires it, parameters are smoothed first. Then a set of contour facts is generated, for example:

```
param(Utterance, Frame_Nr, Contour).
```

These contour facts are used by the rules, to search for (combinations of) events occurring in (combinations of) parameters. If values are needed, we use them relative to the minimum and maximum values of the current utterance, thus avoiding hard-coded thresholds. As our primary concern is not speed but knowledge, the recognition process can take a while. However, this problem is not caused by Prolog, backtracking through the rules and its internal database, but is mainly due to the interface between Prolog and the database. As an example we present the rules that have been developed for the recognition of voiceless plosives (**LP**) and voiceless fricatives (**LF**). Below, we put them into words.

LP1: Look for a deep *valley* in parameter `lpc gain`, for which the steepness of the lines to the first *peaks* forwards and backwards is greater than two.
LP2: Look for a *valley*, not too high, in the combined three highest `frequency bands` (3-5 kHz), with the preceding *peak* being not too high. Near this point also the parameters `zero crossing rate` (`zcr`), `energy` and `lpc gain` should have a *valley*, the one in `lpc gain` being very deep.
LP3: Look for a very deep *valley* in the smoothed parameter `lpc gain`, followed by a *rise* in the smoothed parameters `lpc gain`, `zero crossing rate`, `energy` and the combined three highest `frequency bands`. The relative value of the `zcr` should not be too small at this point.
LF1: Look for a `pitchless` segment in which the highest `frequency band` and `zero crossing rate` have high relative values.

In the LP-rules, pitch is used to calculate measures of certainty. The parameter `lpc gain` appears to contain good indications for LP: a deep *valley* or a steep *rise*. To date, we haven't been able to complete the rules for LF. The present rule finds fricatives like /f/ and /ʃ/, but it is very hard to formulate rules that also find /x/ and /χ/, sounds that are frequent in Dutch, with pronunciations that can vary considerably among speakers.