ISCA Archive
http://www.isca-speech.org/archive

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Contribution of Neural Networks for Phoneme Identification in the APHODEX Expert System

Dominique François and Dominique Fohr

CRIN-CNRS & INRIA Lorraine
B.P. 239 F54506 Vandoeuvre-les-Nancy CEDEX

Email: Dominique.Francois@loria.fr Dominique.Fohr@loria.fr

10.21437/ICSLP.1992-379

ABSTRACT

We propose in this study a way to integrate neural networks by an acoustic-phonetic decoder expert system. The goal of this coupling is to improve the recognition rate of plosive and fricative consonants. We first show that the work we carried out aimed at improving the efficiency of the phonetic knowledge base. In a second step we present a new method based of multilayer perceptrons. This is being used to recognise plosives or fricatives and to be coupled to the expert system. We conclude by discussing how to make this cooperation possible in a hybrid system.

1. INTRODUCTION

The goal of our works is to improve French continuous speech acoustic-phonetic decoding. The first realization was a speaker independent expert system. It aims to obtain a phonetic lattice from the speech signal by an automatic decoding. The performances of this system were not as good as those of the human expert, as a result we started a second study to developpe new algorithms of cue extraction and acquire new knowledge about these cues. This seemed to lead to better results. As a matter of fact we know that neural networks offer a good discrimination between phonemes. This kind of methods can thus help our knowledge-based system in the case of a choice between 2 or 3 phonemes.

2. APHODEX

2.1. Introduction

APHODEX is a project which started in 1986 and which aims to use the knowledge of an experienced spectrogram reader to improve acoustic-phonetic decoding. Therefore, we engaged ourselves in the analysis and formalization of the proficiency acquired by phoneticians in continuous speech spectrogram reading. The result of this study has been implemented in the form of an expert system called APHODEX

architecture:

APHODEX is an expert system composed of:

- a preprocessing module to perform a coarse segmentation of the speech signal into macro classes (fricatives, vowels, plosives and sonorants)
- a set of feature extraction procedures, such as formant tracking,burst finding and analysing, friction detectors...)
- a set of production rules
- an inference engine

Two kinds of production rules have been used to formalize the knowledge of the expert: action rules and deduction rules. Action rules can trigger procedures (modification of the segmentation or lattice). The conclusion of deduction rules is a list of phonemes weighted by confidence coefficients. Almost all rules are context-dependent regarding the right and left regarding phonemes. The results is a phoenetic lattice in which each node contains a list of weighted phonemes.

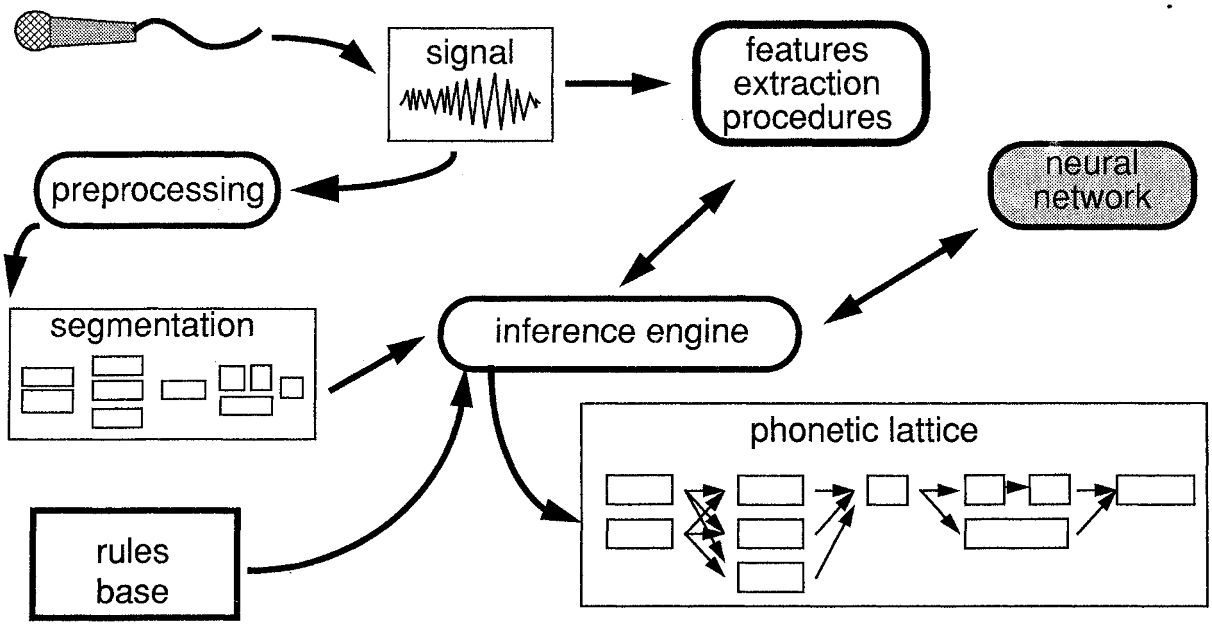
3. CORPORA

For all our experimentations we have used two continuous speech corpora:

<u>- A french corpus made up of:</u>

-The BDSONS corpus «La bise et le soleil». It is made of texts read by 8 speakers.

- A corpus which contains 4 repetitions of 17 sentences said by 18 speakers. For neural networks experimentations 9 speakers were used for training and 9 others for the testing phase.



APHODEX architecture

- An english corpus :

The DARPA TIMIT corpus contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers. The TIMIT database has only been used for neural network's experimentations. 424 speakers were used for the network's training and 176 speakers were used for testing purposes.

4. PLOSIVE IDENTIFICATION

The plosive identification beems to be a difficult acoustic analysis. The task seems to be accomplished quite easily by the expert, but the cues he detects in the spectrogram require very precise algorithms.

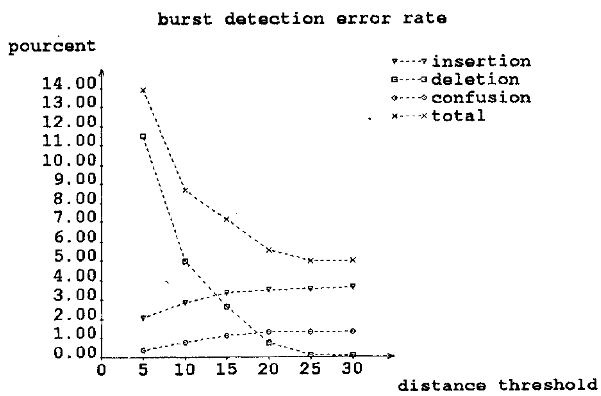
The two regions which contain the right information are the burst, in the plosive, and the first part of the next phoneme. The second region is useful when the phoneme is a vowel because the formant transitions are different according to the preceding plosive. The first region start at the transcient time of the onset, which contains a lot of cues, and end after a more or less long friction period.

Considering that some cues are extracted from an extremely short time period the localisation of this period is of the atmost importance for the rest of the identification. We need a new localization procedure, the result of which have to be reliable and very accurate.

4.1. Burst localization

The principle of the localization is based on computation of a distance between a spectrogram frame and a average burst form. The average form of a burst onset consists of 56 values, 8 frequency band computed during 7 frames. The frequency band are 1000 Hz wide and each value is a 256 order FFT computed on a 392 ms window.

For each spectrogram frame, shifted 392 ms by 392 ms, 8 distances are performed, one per frequency band. We also obtain 8 distance curves in which we can observe minima for each visible burst. We use an experimental adjust threshold in order to determinate the minima which correspond to a zone which may be detected as a burst. The following figure shows the different error rates in several cases of threshold.



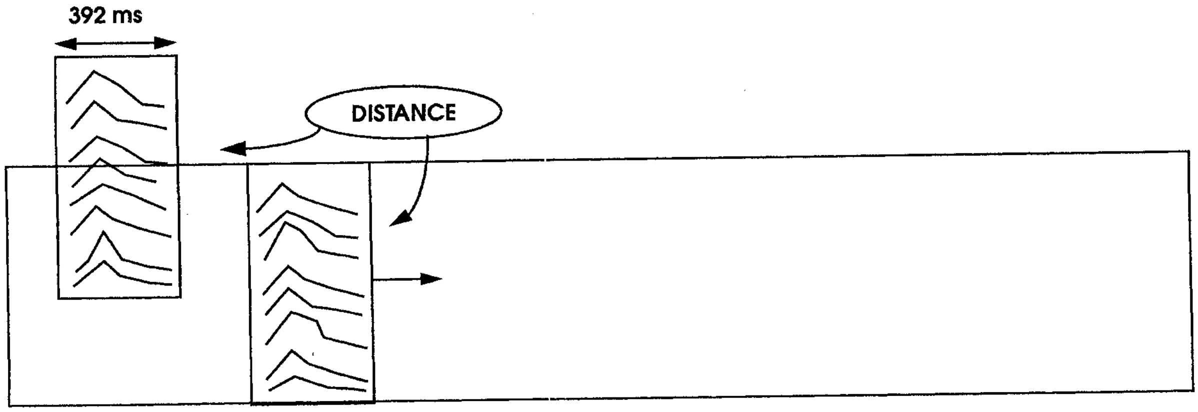
All the plosive for which no burst is visible are rejected. The others are analysed whith a 64 order FFT.

The results are :

good localization rate: 95 %

The error rates are:

- insertion 3.58 %
- deletion 0.06 %
-localization error 1.36 %



4.2. Production rules

New algorithms are used in the knowledge-based system to extract relevant cues from the signal. The cues we use to identify the plosive are :

- the frequency at which energy reaches a maximum. We obtain a list of frequencies from the burst spectrum curve.
- intensity. The maximum of spectrum's energy is used in some rules.
- global shape. The compactness of the spectrum curve is discriminant, as well as its slape.
- energy band ratio. They are useful in describing the contrast between the high energy area and the low energy area in the burst spectrum.

Other cues are used for silences and discriminating voiced and unvoiced plosives. In this case the pitch estimation is not reliable enough, we use the presence of energy in low frequencies.

The following confusion matrix has been obtained with the french corpus. Only the best proposition of the expert system for each phoneme is taken into account.

4.3. Artificial neural networks

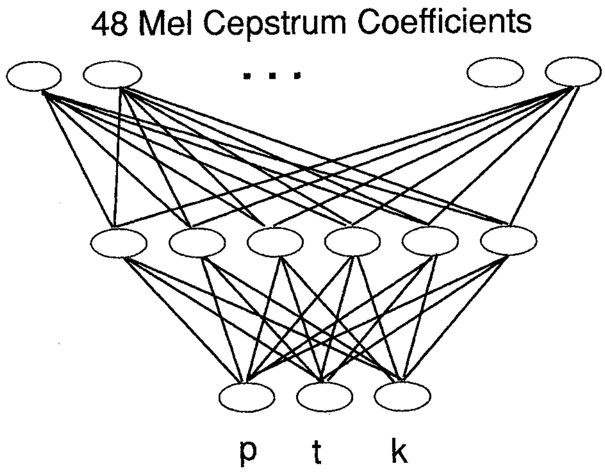
When we ask a phonetician to decode spectrograms, he can tell us which area (part) of the spectrogramm he is looking at. For instance for plosives, he focuses on burst and transitions. We thus try to select (locate) one or two frames in this aeria and then our aim is to use the ability (capacity) of discrimination of ANN to perform recognition.

Architecture of ANN and parametrization [1]

After the selection of the disciminative frames, we compute 12th order Mel Cepstral coefficients on a 32 msec window. The resulting vector is given to a MLP:

- input layer :12 or 24 neurons (one for each MFCC coefficient),
- hidden layer: 5 neurons
- output layer one neuron for each possible phoneme .

We can see an exemple of architecture below:



	Number	p	t	k	b	d	g	Deletion	%
p	498	377	29		6	58	13	12	75.7
t	889	55	503	34	201	10	6	66	56.6
k	560	34	49	308	89	12		55	55.0
b	471	4	16	6	367	10	2	26	77.9
d	988	93	12	1	156	459	7	153	46.5
g	387	33	14		42	15	151	81	39.0
Insertion		40	9		21	5			

The training is done by retropogation of error.

Discrimination tests have been performed for phonemes with suficient occurences in the french corpora. The results are presented below:

context	plosives	occurence s	recognition
/o/ /u/ /)/	/p/, /t/, /k/	144	95 %
	/b/, /d/, /g/	95	92 %
/i/	/p/, /t/, /k/	74	81 %
	/b/, /d/	44	97 %
/e/ /ε/	/p/, /t/	82	90 %
	/b/, /d/	65	95 %
/ə/	/p/, /t/, /k/	55	93 %
/a/	/b/, /d/, /g/	105	90 %
/l/	/p/, /k/	44	90 %
/R/	/p/, /t/	28	90 %

5. FRICATIVE IDENTIFICATION

5.1. Expert system

There are two major groups of production rules for the fricative identification:

- rules using the lower frequency of fricative noise.
- rules using the gravity center of fricative noise.

The results give by expert system are presented in the following table. The occurences of /v/ fricatives in our french corpus is too small and we can not take into account the results concerning that phoneme. In the second colomn we showed the results obtained with only the first proposition of the expert system and in the third colomn results obtained with the two first propositions.

	The first	The 2 firsts
v	20.5 %	45.6
ch	79.2	79.2
gh	51.2	71.8
s	77.7	77.7
z	72.1	80.5

Tableau 1 :

5.2. Neural network

The neural network results are presented in the following tables. The first one was obtained with the french corpus, the second one with the TIMIT database.

Tableau 2 :

fricatives	identification rate
/f/ /s/ /ch/	92 %
/v/ /z/ /gh/	82 %
/s/ /ch/	97 %
/v/ /gh/	96 %
/v/ /z/	91 %
/z/ /gh/	93 %

Tableau 3 :

fricatives	identification rate
/f/ /s/ /ch/	89 %
/s/ /ch/	92 %

6. CONCLUSION

The original idea of this work is that APHODEX is now using a new method which is not an analytic one. The advantage of using neural networks is that it do not need knowledge acquisition, the perceptron "learns" automatically. The first difficulty of expert systems is indeed the long phase of perfecting in colaboration whith the human expert.

BIBLIOGRAPHY

[1] Y. Gong and J.P. Haton "Towards a General Signal Interpretation System Signal-to-symbol Conversion Level" IEEE ICPR 1990 p 79-84 Atlantic City USA

[2] D. François and D. Fohr, "Première évaluation d'APHODEX, système expert pour le décodage acoustico-phonétique de la parole continue", JEP 1990, Montréal Canada

[3] F. Lonchamp "Reading Spectrograms : the View of the Expert" in "Fundamentals in Computer Understanding : Speech and Vissio" , J.P. Haton Editor Cambridge University Press 1987.

[4] N. Carbonell, D. Fohr and J.P. Haton "APHODEX, an Acoustic-Phonetic Decoding Expert System" International Journal of Pattern Recognition and Artificial Inteligence vol 1 No2 pp. 207-222 1987

[5] D. Memmi, M. Eskenazi, J Mariani, A. Nguyen-Xuan "Un système expert pour la lecture de sonagrammes", Speech Com. vol 2, No 2-3, 1983 pp 234-236