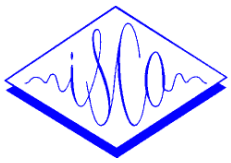

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

EXPERIMENTS IN THE USE OF AN AUTOMATIC LEARNING SYSTEM FOR ACOUSTIC-PHONETIC DECODING

C. Montacié [1-2], M.-J. Caraty [1] & X. Rodet [1]

[1] **Laforia** Université Paris 6, CNRS URA 1095 - 4, place Jussieu 75252 Paris Cedex 05 France
[2] **Télécom Paris** Dépt SIG, CNRS URA 820 - 46, rue Barrault 75634 Paris Cedex 13 France

ABSTRACT

Results are reported of experiments in the use of Charade, an Automatic Learning System, to classify phonetic macro-classes. A preliminary evaluation of the Charade system is carried out on a reference database of continuous speech and compared to an usual classifier (i.e., Hamming Distance Nearest Neighbor) and a neural net based technique (i.e., Modified Hopfied Net). Preliminary results of classification of phonetic macro-classes can be summarized as follows :
— For a given reasonable error rate, Charade classifier gives the lowest rejection rate.
— An important advantage of Charade lies in the ability to analyse and interpret the production rules. For instance, a rule can be interpreted in terms of cues relevant to features. Spectral masks of vowel, drawn from analysis of the most frequent clustering rules found for each of them, are shown to be coherent with phonetic knowledge.

1. INTRODUCTION

Knowledge acquisition based on observation or experience is of major necessity as far as the performance of Automatic Recognition Systems is concerned. An automatic technique for generalization and learning of a production rule system from a training set of examples should allow the acquisition and the constitution of a consistent knowledge database.

The Charade system [1] was designed to detect logical or statistical regularities existing in a set of examples and to generate a production rule system reflecting such regularities. The learning technique is mainly characterized by a principle of representation based on the Hilbert Cube taking advantage of mathematical properties and optimized exploration of the description space for the rules generation.

The results of the Charade automatic learning (Ch) will be presented for the classification of phonetic macro-classes on a reference database of continuous speech. As points of comparison, we chose two different classification techniques: Hamming Distance Nearest Neighbor (HDNN) [2] and a Modified Hopfield Net (MHN) [3].

2. CHARADE : AUTOMATIC LEARNING OF A PRODUCTION RULE SYSTEM

Charade stands, in french, for Hilbert Cubes Applied to Representation and Learning Based on Examples Description. From a description language, a set of axioms reflecting the language semantics and a set of examples expressed in that language, Charade generates a consistent production rule system.

2.1. Description of examples

The description d(E) of an example E is a conjunction of descriptors :

$$d(E) = d_1 \,\&\, d_2 \,\&\, \ldots \,\&\, d_D$$

Each descriptor d_i being an atomic proposition or the negation of an atomic proposition. With such a description of examples, the elementary operations in the field of learning (i.e., generalization and discrimination) are computed as simple conjunction of descriptors. For instance, the least generalization of two examples is the conjunction of common descriptors of these examples.

2.2. Representation space : the Hilbert Cube

The representation space of the learning set (e.g., a set of n examples) is a Hilbert Cube : a unitary hypercube in an n dimensional space. In this representation space :
— Each axis is associated to an individual.
— Each vertex corresponds to a subset of individuals characterized by its least generalization.
— The belonging of an individual to a vertex is defined by the projection of the vertex onto the corresponding axis.
— Edges of the cube are considered as inheritance relationships towards the origin allowing an optimization of the representation. Indeed, a vertex inherits of every vertex above it in the hierarchy of the cube.

2.3. Principles of logical rule generation

The induction principle is the following : given two descriptors d_1 and d_2 then if all the examples of the learning set containing d_1 contain also d_2 then $d_1 \Rightarrow d_2$. To introduce the generation principle, let us consider the two following representation spaces :
— The Hilbert Cube of Examples (C_{Ex}), describing the set of subsets of the learning examples, ordered by set inclusion.
— The Hilbert Cube of Descriptors (C_{Des}), describing the set of descriptor conjonctions, ordered by logical implications.
Then, the principle for the generation of the rules will consist to define the link between these two ordering relations by an exploration of the Cube of Descriptors

For the generation of a potential rule, four functions are defined as follows :
— $\delta : C_{Ex} \longrightarrow C_{Des}$
For each vertex of C_{Ex} (i.e., a set of examples),

δ associates the vertex of C_{Des} corresponding to the least generalization of the set.

— $\gamma : C_{Des} \rightarrow C_{Des}$
For each vertex of C_{Des} (i.e., a descriptors conjunction), γ associates the vertex of C_{Des} corresponding to the set of all the examples for which this descriptors conjunction appears.

The application $\beta = \delta \circ \gamma : C_{Des} \rightarrow C_{Des}$, has the property to show up all the logical relations present in the learning set.

— ω and τ are defined to eliminate redundancies related to inheritance relationship and implication transitivity.

With the previous function definitions we can obtain for each vertex V_{Des} of the Cube of Descriptors, a potential rule of the type : $V_{Des} \Rightarrow \tau(\omega(\beta(V_{Des})))$. For the generation of the complete rule system, an exhaustive exploration of the Cube of Descriptors is excluded. For feasability of the technique of generation, various limitation theorems for the exploration are used.

2.4. Limitation theorems

Two basic theorems determine nullity criteria of vertices of the Cube of Descriptors (i.e., vertices for which $\tau \circ \omega \circ \beta$ is empty). The goal is to eliminate irrelevant vertices (i.e., the vertices who will not generate new rules).

Other theorems are introduced as constraints. They are related to the desired properties of the rule-based system. For instance, the descriptors structuration, the noise factor, the terminal condition of the rules, etc... Fifteen parameters may be adjusted in the Charade system.

3. APPLICATION FOR SPEECH PROCESSING

The application in the field of speech, consists in the classification of spectra according to phonetic macro-classes. Short time spectrum is an example in our experiment. Spectra are manually labeled from a continuous speech data base. The description of an example is a conjunction of binary descriptors.

3.1. Parametrisation

The selection of a parametric representation of acoustic data is an important task in the design of any speech experiment. The usual objective is to eliminate information not pertinent with regard to phonetic analysis and to enhance the aspects of signals which contribute to phonetic differences. Knowledge of audition leads us to look after the frequency content of acoustic signals and more precisely to look for spectral masses.

For our experiment, we choose a binary description of spectral power balances. For each labeled phone, a signal frame of 32 ms duration is extracted. Channels on the power spectrum $\{E_k\}$ are computed through a triangular filter according to Mel frequency scale [4]. Four analysis levels are considered. At level i, the frequency range is divided into 2^i channels. For any pair of contiguous channels (e.g., k and k+1) a descriptor is computed as the value of the logical expression $(E_k > E_{k+1})$ (cf. table 4). If frequency channels are narrow enough, such a parametrization allows for localisation of spectral masses such as formants.

4. EXPERIMENTS

Concerning the classification of phonetic macro-classes, four oppositions are studied :

a- Orals versus Nasal Vowels
(i.e., [i, e, E, A, o, O, y, u, eu] vs [an, on, un])

b- Voiced versus Unvoiced Fricatives
(i.e., [v, z, j] vs [f, s, ch])

c- Voiced versus Unvoiced Plosives
(i.e., [b, d, g] vs [p, t, k])

d- Glides versus Nasal Consonants
(i.e., [l, r] vs [m, n])

The various classification techniques we tested (i.e., HDNN, MHN and Ch) were carried out on a reference database.

4.1. Reference speech database

The reference database we used is the Acoustic Corpus SYL which is a part of BDSONS supported by GRECO *Communication parlée*, CNRS, France. This corpus contains 192 diphone dedicated sentences of continuous speech per speaker.

4.2. Reference systems

In order to evaluate the learning and the accuracy rate of the Charade system, we chose two reference systems : Hamming Distance Nearest Neighbor and a Modified Hopfield Net.

The HDNN classifier is based on the Hamming distance. This distance is the number of bits in the test which does not match the corresponding reference bits.

Artificial neural net models have been studied for many years in hope of achieving human-like performance in the fields of speech and image recognition. These models are composed of many non-linear computational elements (i.e., nodes) operating in parallel. The Classical Hopfield Net can be described as follows :

— The net is defined by N nodes containing hard limiter and binary input and output $\mu_i(t)$. N = D + E : D nodes represent the D descriptors of the parametrisation, E nodes coding the various classes are used for the classification.

— Each node is fed back to all other nodes via weights t_{ij}
The learning weights are computed from M patterns $\{x^S\}$, (S = 1,..., M) with the following formula :

$$t_{ij} = \begin{cases} \sum_{s=1}^{M} (x_i^S * x_j^S) & \text{for } 1 \le i \le N,\ 1 \le j \le N,\ i \ne j, \\ 0 & \text{for } i = j. \end{cases}$$

— The value of a node $\mu_i(t+1)$ is computed as follows via a hard limiter F

$$\mu_i(t+1) = F\left(\sum_{j=1}^{N} t_{ij} * \mu_i(t)\right) \quad \text{for } 1 \le i \le N,$$

with $F(x) = -1$ if $x < 0$, $+1$ if $x > 0$.

The Classical Hopfield Net has major limitation when used as a classifier. Hopfield showed that the net performances decrease when number of classes (E) is less than 0.15 the number of nodes (N) in the net. In our experiment, the results obtained with this algorithm were

quite insufficient (i.e., a recognition rate lower than 50%). Thus, we modified this algorithm.

— The **first modification** concerns the offset S_i of the hard limiter F :

$$F(x) = \begin{cases} -1 \text{ if } x < S_i, \\ +1 \text{ if } x > S_i. \end{cases}$$

S_i is defined as the optimal offset node. From training patterns, for every node μ_i we calculate mean and standard deviation of positive and negative activations (i.e., m_{pi}, d_{pi}, m_{ni}, d_{ni}). With such notations, the offset is computed with this formula :

$$S_i = m_{ni} + d_{ni} * \frac{m_{pi} - m_{ni}}{d_{ni} + d_{pi}}$$

— The **second modification** concerns the decision principle : after convergence a supplementary step is computed without hard limiter . Decision is taken with the optimal activation defined below :

$$((\sum_{j=1}^{N} t_{ij} * \mu_i(t)) - S_i) / d_{pi}$$

This decision principle is equivalent to an orthogonalization procedure [5]. In our experiment, these modifications lead to a high improvement of the accuracy rate.

5. RESULTS

For each experiment :
— The training set, the test set and the rules number are showed on the table 1.
— Three results areas (i.e., accuracy, error and rejection) corresponding to the various classifier tested (i.e., HDNN, MHN and Ch) are showed on the table 2 on the following page. To take into account the bias of the test set (i.e., phonemes occurences), the results are computed as means corresponding to the various phonetic classes of the test set.

oppositions	Training Set	Test Set	Rules Number
Oral Vowels	54	1209	13
Nasal Vowels	54	294	15
Voiced Fricatives	54	192	3
Unvoiced Fricatives	54	219	2
Voiced Plosives	54	232	7
Unvoiced Plosives	54	231	4
Glides	54	264	6
Nasal Consonants	54	145	9

Table 1.
Learning Characteristics

5.1. Results Analysis

HDNN and MHN, used without any efficient rejection criteria, give the best accuracy rate and perform alike. However, their error rates (i.e., about 12%) are rather high. For acoustic-phonetic decoding, the goal is mainly to minimize error rates. Given a same reasonable error rate (e.g., 5%), an experiment shows Charade gives the highest accuracy rate (cf. table 3). This experiment consisted in using tresholds for rejection criteria (i.e., maximum distance for HDNN and minimum activation for MHN).

Classifier & tresholds	accuracy rate	error rate	rejection rate
Ch	**68.8**	**6.6**	**23.6**
HDNN (5)	76.6	9.9	13.5
HDNN (4)	**57.7**	**5.9**	**36.4**
HDNN (3)	28.1	1.8	70.1
MHN (2)	81.4	7.3	11.3
MHN (3.5)	**65.0**	**5.9**	**28.1**
MHN (3.75)	58.2	4.4	37.4

Table 3.
Classifier results for HDNN and MHN with different tresholds for rejection criteria

5.2. Rule System Analysis

An important advantage of the Charade system lies in the ability to analyse and interpret the generated rules.

During the test set classification step the analysis of rules activation frequency, per macro-classes, shows up the significant clustering rules. For instance, for Oral versus Nasal Vowels classification and with the indication of the most concluded phonemes, the found significant rules are the following :

— 4 principal rules conclude on Oral Vowels :

R1. E[580-935]<E[935-1390] & E[1390- 2025]>E[2025-2950]
=> [A, E, e, y]
R2. E[62-312]>E[312-580] & E[580-935]<E[935-1390]
=> [e, i, y]
R3. E[62-580]<E[580-1390] & E[1390- 2025]>E[2025-2950]
=> [E, A, e, y]
R4. E[62-580]<E[580-1390] & E[62-312] >E[312-580]
=> [e,y]

— 2 principal rules conclude on Nasal Vowels :

R1. E[312-580]<E[580-935] & E[1390-2025]<E[2025-2950]
=> [an,un]
R2. E[580-1390]>E[1390-2950] & E[312-437]<E[437-580]
& E[748-935]>E[935-1142]
=> [an]

The rules can be interpreted in terms of cues (e.g., invariant primitives) relevant to features (e.g., Oral versus Nasal).

Spectral masks of vowels, drawn from the same analysis carried out per phoneme, are shown to be coherent with phonetic knowledge. Examples of two well classified cardinal vowels (e.g., [A, i] are shown (cf. Fig 1 & 2)

6. CONCLUSIONS

Charade evaluates an information measure of the parametrisation in terms of invariance. When experiments are carried out on the whole group of macro-classes, Charade does not work efficiently. A reason is the result of rule system autocoherence : generated rules cannot even cover 50% of the training set. This point is very important as it puts in question the validity of the description language for the considered problem. It is expected that the performance will improve with a better description language.

An advantage of the Charade system is also to be a principle of decision allowing the integration of heterogenous data (e.g., informations coming for various signal processing methods). We decribed and tested the generation of logical rules. Charade allows the generation of statistical rules to take

into account mistaken example description and phonetic classes recovering. Such rules will be tested in order to integrate two speech processing methods (i.e., Temporal Decomposition and Rupture Models) [6][7][8].

Acknowledgment : we are especially indebted to Jean-Gabriel Ganascia for providing us with the CHARADE system and for making helpful suggestions.

REFERENCES

[1] J.G. GANASCIA : AGAPE et CHARADE : **deux techniques d'apprentissage appliquées à la construction de bases de connaissance.** Thèse d'état, Université Paris-Sud, 1987.

[2] R.G. GALLAGER : **Information Theory and Reliable Communication.** John Wiley & Sons, New York, 1968.

[3] J.J. HOPFIELD : **Neural Networks and Physical Systems With the Emergent Collective Computational Abilities.** Proc Nat. Acad. Sci. USA vol 79, pp 2554-2558, 1982.

[4] DAVIS & MERMELSTEIN : **Comparaison of Parametric Representations for Monosyllabic Word Recognition in Continuous Spoken Sentences.** IEEE-ASSP. 28.4, 1980.

[5] D.J. WALLACE : **Memory and Learning in a Class of Neural Models.** Workshop on Lattice Gauge Theory, Wuppertal, 1985

[6] R. ANDRE-OBRECHT : **A New Statistical Approach for the Automatic Segmentation of Continuous Speech.** IEEE Trans. ASSP, vol. 36, pp. 29-40, 1988.

[7] F. BIMBOT, G. CHOLLET, P. DELEGLISE & C. MONTACIE : **Temporal Decomposition and Acoustic-Phonetic Decoding of Speech.** Proc ICASSP-88, pp. 425-428.

[8] A.M.L. Van DIJK-KAPPERS & S.M. MARCUS : **Temporal Decomposition of Speech.** Speech communication, Vol. 8, No 2. 1989.

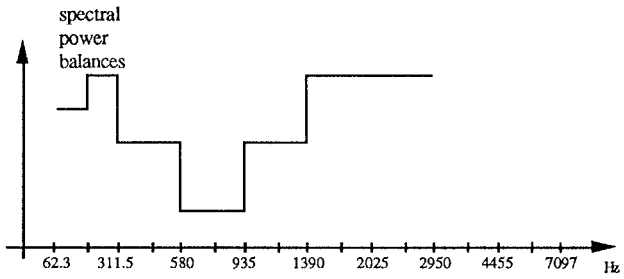

Figure 1
Spectral mask for the vowel [i]

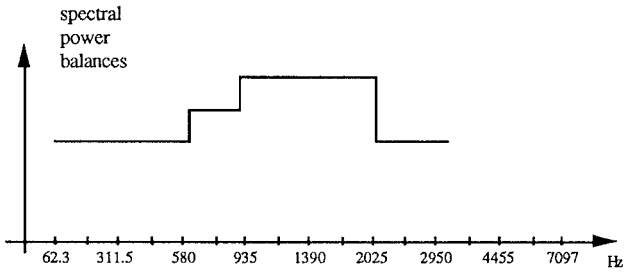

Figure 2
Spectral mask for the vowel [a]

Macro-classes	accuracy rate			error rate			rejection rate		
	HDNN	MHN	Ch	HDNN	MHN	Ch	HDNN	MHN	Ch
Oral Vowels	78.5	87.7	67.9	21.5	12.3	4.8	0.0	0.0	27.3
Nasal Vowels	83.7	67.8	57.4	16.3	32.2	8.7	0.0	0.0	33.7
Mean	*79.8*	*82.7*	*65.3*	*20.2*	*17.3*	*5.8*	*0.0*	*0.0*	*18.9*
Voiced Fricatives	96.3	99.3	77.1	3.7	0.7	8.3	0.0	0.0	14.6
Unvoiced Fricatives	99.4	88.9	92.7	0.6	11.1	0.0	0.0	0.0	7.3
Mean	*97.8*	*94.1*	*84.6*	*2.2*	*5.9*	*4.1*	*0.0*	*0.0*	*11.3*
Voiced Plosives	87.1	85.6	77.1	12.9	4.4	4.3	0.0	0.0	18.6
Unvoiced Plosives	95.2	98.1	43.4	4.8	1.9	14.3	0.0	0.0	42.3
Mean	*91.1*	*91.8*	*60.3*	*8.9*	*8.2*	*9.3*	*0.0*	*0.0*	*30.4*
Glides	87.3	97.1	54.5	12.7	2.9	9.9	0.0	0.0	36.6
Nasal Consonants	96.4	81.5	81.1	3.6	18.5	7.5	0.0	0.0	11.4
Mean	*91.8*	*89.2*	*67.8*	*8.2*	*10.8*	*8.7*	*0.0*	*0.0*	*33.5*
Mean	**87.8**	**88.0**	**68.8**	**12.2**	**12.0**	**6.6**	**0.0**	**0.0**	**24.6**

Table 2.
Classification results for HDNN, MHN and Ch

	Niveau 1		
d1	E[62.3 - 1390]	>	E[1390 - 7097]
	Niveau 2		
d2	E[62.3 - 580]	>	E[580 - 1390]
d3	E[580 - 1390]	>	E[1390 - 2950]
d4	E[1390 - 2950]	>	E[2950 - 7097]
	Niveau 3		
d5	E[62.3 - 311.5]	>	E[311.5 - 580]
d6	E[311.5 - 580]	>	E[580 - 935]
d7	E[580 - 935]	>	E[935 - 1390]
d8	E[935 - 1390]	>	E[1390 - 2025]
d9	E[1390 - 2025]	>	E[2025 - 2950]
d10	E[2025 - 2950]	>	E[2950 - 4455]
d11	E[2950 - 4455]	>	E[4455 - 7097]
	Niveau 4		
d12	E[62.5 - 186.9]	>	E[186.9 - 311.5]
d13	E[186.9 - 311.5]	>	E[311.5 - 436.7]
d14	E[311.5 - 436.7]	>	E[436.7 - 580]
d15	E[436.7 - 580]	>	E[580 - 748]
d16	E[580 - 748]	>	E[748 - 935]
d17	E[748 - 935]	>	E[935 - 1142]
d18	E[935 - 1142]	>	E[1142 - 1390]
d19	E[1142 - 1390]	>	E[1390 - 1678]
d20	E[1390 - 1678]	>	E[1678 - 2025]
d21	E[1678 - 2025]	>	E[2025 - 2439]
d22	E[2025 - 2439]	>	E[2439 - 2950]
d23	E[2439 - 2950]	>	E[2950 - 3595]
d24	E[2950 - 3595]	>	E[3595 - 4455]
d25	E[3595 - 4455]	>	E[4455 - 5622]
d26	E[4455 - 5622]	>	E[5622 - 7097]

Table 4.
26 binary descriptors corresponding to 4 analysis levels