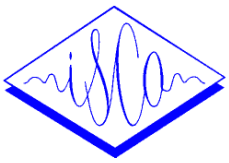

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

INTEGRATION OF FREQUENTIAL AND TEMPORAL STRUCTURATIONS IN A SYMBOLIC LEARNING SYSTEM

Marie-José Caraty, Claude Montacié & Claude Barras

LAFORIA - Université Paris 6, CNRS-URA 1095 - 4, place Jussieu - 75252 Paris Cedex 5 - France

ABSTRACT

This paper presents an approach to integrate frequential and temporal structurations of the speech signal in a symbolic learning and rule-based recognition process. This approach is evaluated on experiments of vowels recognition in continuous speech and compared with a statistical approach. To take into account the frequential and temporal structurations, we choose the Clustering of Spectral Peaks (CSP) measure based on peak parameters space and the Generalized Temporal Decomposition (GTD) modeling the spectral evolution. Charade is the chosen symbolic learning system. From our experiments, a priori strategic temporal information, such as localization or segmentation obtained from GTD technique using CSP measure, is not shown to be useful information for vowels identification. Our comparative study symbolic versus statistical approach is very encouraging for further research on symbolic process. Indeed, the results obtained from symbolic approach are quite comparable to the best results obtained from statistical approach.

10.21437/ICSLP.1992-118

I. INTRODUCTION

We propose an approach to integrate frequential and temporal structurations of the speech signal in a symbolic learning and rule-based recognition process. Charade [6] is an automatic technique of generalization from observations. This system is designed to detect logical regularities from a set of examples expressed in a description language and to generate the production rule system reflecting those regularities. In a previous study [10], Charade's capability to find discriminant rules on phonetic macro-classes was tested. To now improve the signal processing we choose to take into account the temporal evolution of the speech signal. To this end, we select an original model of the spectral evolution : the Generalized Temporal Decomposition [9]. This method analyses the deviations of the spectral trajectory to localize the main events of the speech signal. The distortion measure used to represent the spectral trajectory is a perceptive-based measure : the Clustering of Spectral Peaks measure [4]. The cooperation of Charade system and the GTD technique using the CSP measure allows the integration of three a priori significant structurations in the applications of Speech Recognition : the frequential structuration (i.e., the measure based on the clustering of spectral peaks), the temporal structuration (i.e., the targets and the interpolation functions reflecting the acoustic events) and the logical structuration (i.e., the learned production rule system).

The evaluation of this approach consists in a series of experiments on vowels recognition. For these experiments, various representation spaces of the vowels are tested. Some representations use the localization or the segmentation issued from GTD technique and CSP measure. Finally, the symbolic approach is compared to a statistical approach on similar experiments. The identification principle we choose is the nearest-neighbor in the meaning of LPC-based measures. The selected measures are the following : euclidean metric on Log Area Ratio Coefficients and Linear Prediction Cepstrum Coefficients, Log Likelihood Ratio measure and Clustering of Spectral Peaks measure.

II. FREQUENTIAL STRUCTURATION : CLUSTERING OF SPECTRAL PEAKS MEASURE

As far as vowels recognition is concerned, our choice of the parameter space is justified by the well-known importance of the formants for the characterization, the discrimination and the perception of the vowels. The chosen spectral representation is based, by analogy with formants, on the characteristics of the spectral maxima/peaks. In this representation space, the inter-spectra distortion measure used is the Clustering of Spectral Peaks measure (CSP). This measure is based on perceptive criteria of the sound perception [4].

2.1. Spectral peaks parameter space

A short-time spectrum S is represented by the set of its spectral peaks (i.e., local maxima) $\{P_k\}_{(k=1,..,K)}$ characterized by their central frequency F_k (Hz), their 3 dB bandwith B_k (Hz) and their intensity I_k (dB) :

$$S = \{P_k(F_k, B_k, I_k)\}_{(k=1,..,K)}$$

The signal is analysed by a 16th order linear prediction over 25.6 ms windows. A robust algorithm [5] is used for the computation of the peak parameters from the LPC spectrum. Indeed, the robustness is indispensable for the use of the Generalized Temporal Decomposition technique.

2.2. Clustering of Spectral Peaks measure

The original Clustering of Spectral Peaks measure [4] has proved to have many advantages [4] [12] [13]. Its computation is summarized as follows.

2.2.1. Inter-peaks distortion measure

A local inter-peaks measure $\delta_{CSP}(P_i, P_j)$, between two peaks of distinct spectra, is introduced for the measurement of a local spectral distortion and computed with the following formula :

$$\delta_{CSP}(P_i,P_j) = \omega_F(F_i).e_F(P_i,P_j)+\omega_B(F_i).e_B(P_i,P_j)+\omega_I(F_i).e_I(P_i,P_j)$$

$$e_F(P_i,P_j) = \frac{|F_i-F_j|}{F_i+F_j}; \qquad e_B(P_i,P_j) = \frac{|B_i-B_j|}{B_i+B_j}; \qquad e_I(P_i,P_j) = |I_i-I_j|$$

The weighting tables on frequency $\{\omega_F(F_i)\}$, bandwith $\{\omega_B(F_i)\}$ and intensity $\{\omega_I(F_i)\}$ are computed from the probability distribution functions of the first four formants on the frequency scale $(\{\Pi_{F_1}(F_i)\}, \{\Pi_{F_2}(F_i)\}, \{\Pi_{F_3}(F_i)\}, \{\Pi_{F_4}(F_i)\})$ considered [3] and from type values of weighting coefficients for the first formants $\{\Omega_F(\mathcal{F}_i), \Omega_B(\mathcal{F}_i), \Omega_I(\mathcal{F}_i)\}_{(i=1,..,4)}$ [4].

2.2.2. Inter-spectra distortion measure

The global inter-spectra measure $D_{CSP}(S^R, S^T)$ is computed from the optimal clusterings of peaks of the given spectra S^R and S^T. The clustering of a peak P_i^R of S^R to a peak P_j^T of S^T is measured by the inter-peaks measure $\delta_{CSP}(P_i^R, P_j^T)$ and is defined

optimal when :

$$P_j^T = \text{Argmin}_{\{P^T \in S^T\}} \{\delta_{CSP}(P_i^R, P^T)\}$$

Let Δ be the matrix of the inter-peaks measures of the two given spectra :

$$\Delta = \{\Delta_{ij} = \delta_{CSP}(P_i^R, P_j^T)\}_{(i=1,..,I\ ;\ j=1,..,J)}$$

The global inter-spectra measure $D_{CSP}(S^R, S^T)$ is computed as the average of the distinct optimal clusterings located on the rows and the columns of the matrix Δ.

III. TEMPORAL STRUCTURATION : GENERALIZED TEMPORAL DECOMPOSITION

The classical Temporal Decomposition (TD) [1] describes the spectral evolution of signals. It uses a linear interpolation model and analysis techniques based on a criterion of minimization of quadratic error. Rigorously, the recognition operator must be based on euclidean metric. Usually used for its elementary computation, this distance isn't however the most performant spectral distortion measure. The interesting property of the Generalized Temporal Decomposition (GTD) [9] is to use a linear interpolation model generalized itself to any distortion measure D.

3.1. Temporal Decomposition model

The Temporal Decomposition (TD) [1] describes the spectral evolution of a speech signal, represented by the spectral vectors $\{y_n\}_{(n=1,..,N)}$, by a linear interpolation model :

$$\hat{y}_n = \sum_{i=1}^{q} g_i\, \phi_i(n)$$

The estimation $\hat{y}_n$ of the spectral vector y_n is a linear combination of a limited number of spectral vectors $\{g_i\}_{(i=1,..,q)}$ called spectral targets. To each target g_i is associated a compact interpolation function ϕ_i. The targets $\{g_i\}$ represent the spectral content of the signal, the interpolation function $\phi_i(n)$ represents the influence of the i^{th} target for the estimation of the n^{th} spectral vector y_n. The TD algorithm searches for the optimal choices of the number q of targets, the interpolation functions $\{\phi_i(n)\}$ and spectral targets $\{g_i\}$. A priori, the single constraint applied to the functions ϕ_i : they must be positive and time-limited.

3.2. Generalized Temporal Decomposition model

To generalize the Temporal Decomposition [9], we propose an approximation of the generalized linear interpolation model using the factorial analysis. We project a part of the spectral trajectory in a new space with euclidean distance and we look for the interpolation function representing this portion of signal.

Given the spectral trajectory represented by the vectors $\{y_n\}_{(n=1,..,N)}$ of any metric space using a distance D. The estimation of the interpolation function requires the computation of the first n eigen vectors of the matrix S of dimension NxN whose elements $\{S_{ij}\}_{(i,j=1,..,N)}$ are computed as follows :

$$S_{ij} = -\sum_{k,l=1}^{N} D(y_k, y_l)^2 / 2N^2 - D(y_i, y_j)^2 / 2$$

$$+ (\sum_{k=1}^{N} D(y_i, y_k)^2 + \sum_{k=1}^{N} D(y_k, y_j)^2) / 2N$$

The first n eigen vectors of the matrix S have to represent at least 95 % of the inertia of the trajectory (i.e., 95 % of the trace of S). The chosen interpolation function is a linear combination of these first n eigen vectors. The linear combination has to maximize the correlation between the interpolation function computed and a rectangular characteristic function.

IV. LOGICAL STRUCTURATION : SYMBOLIC LEARNING WITH CHARADE

Charade system [6] 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 and, in terms of control of the combinations, an intelligent exploration of the description space for the rules generation.

4.1. Representation space : the Hilbert Cube

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. The description d(E) of an example E is a conjunction of descriptors : $d(E) = d_1 \wedge d_2 \wedge \ldots \wedge d_D$. Originally, each descriptor d_i is an atomic proposition or the negation of an atomic proposition.

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. The mathematical properties of this representation space allow an optimization of the representation and guarantee the feasibility of the learning technique.

4.2. 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 following two representation spaces :

— The Hilbert Cube of Examples (C_{Ex}), describing the set of subsets of the learning examples, ordered by the set inclusion.

— The Hilbert Cube of Descriptors (C_{Des}), describing the set of descriptor conjonctions, ordered by the logical implications.

Then, the principle for the generation of the rules consists in defining 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 :

— $\delta : C_{Ex} \rightarrow C_{Des}$. 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.

— $\gamma : C_{Des} \rightarrow C_{Ex}$. For each vertex of C_{Des} (i.e., a descriptors conjunction), γ associates the vertex of C_{Ex} 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 that are present in the learning set.

— ω and τ are defined to eliminate redundancies related to inheritance relationships 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 feasibility of the technique of generation, various limitation theorems for the exploration are used. These theorems aim at eliminating irrelevant vertices of the Cube of Descriptors (i.e., the vertices which will not generate new rules) or introduce constraints related to the desired properties of the rule-based system.

4.3. Example description

For our speech processing application, an example is a short-time spectrum represented by the set of its spectral peaks (cf. §2.1.). A previous description [5] of 30 binary descriptors $\{d_k\}_{(k=1,..,30)}$ computed from this representation was tested. Let m and v be the two descriptor values : d_k=m (resp. d_k=v) corresponds to the presence of a spectral masse (resp. valley) in a frequency band B_k (cf. Figure 1). The frequency bands $\{B_k\}_{(k=1,..,30)}$ are 63 Mels wide contiguous bands.

To enhance this description of the absolute position of the spectral peaks, we choose to add information of relative position of the first three peaks (i.e., P_1/P_2, P_1/P_3, P_2/P_3). To measure the relative distribution (e.g., P_i/P_j) we choose to quantify the value $\Delta_{F_{ij}} = |F_i-F_j|/(F_i+F_j)$ by one or more symbols "+" and to represent it at the central frequency $(F_i+F_j)/2$ (cf. Figure 1).

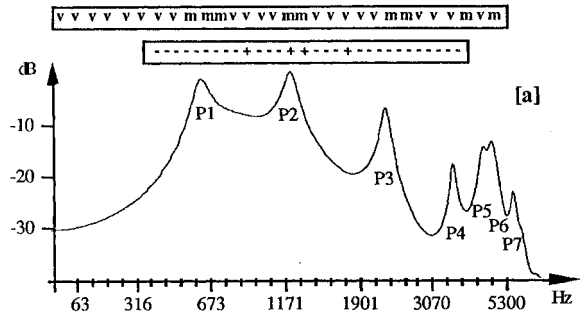

Figure 1. Description of an example

V. EXPERIMENTS

The Charade system is applied to the identification of oral/nasal vowels in a continuous speech signal. The representation of the vocalic entities is problematic. Indeed, these entities are signal continuums : the phonetic realisations of vowels contained in a speech signal. Taking into account locally or globally the structuration of the speech signal, various representations are tested. The learning of vocalic entities with Charade generates a production rule system for the classification of the vowels : each rule concludes on the phonetic label of a vowel. From this rule system, the principle of identification of a test-entity is very simple and rapid. It is based on the number of the rules verifying accurately the test-entity. To compare the obtained results, we choose similar experiments using the nearest-neighbor identification principle with classical LPC-based measures.

5.1. Speech database

For the experiments, the reference database we use is the Acoustic Corpus SYL which is a part of the french database BDSONS supported by GRECO Communication Parlée, CNRS, France. The corpus SYL_y, numbered (y) from 1 to 12, allow the study of 192 diphones $[C_nV_y]$ counted for the oral/nasal vowels $\{[V_y]\}_{(y=1,..,12)}$ and the consonants $\{[C_n]\}_{(n=1,..,16)}$ of French. Each corpus SYL_y contains 16 sentences $\{P_n\}_{(n=1,..,16)}$, each sentence P_n presenting various occurrences of the diphone $[C_nV_y]$. The database contains about 4000 phonemes including 1980 oral/nasal vowels. The speech signals are manually labeled according to the principle of large labeling [2]. Few distinctions of phonetic labels are not taken into account (e.g., {[œ], [ə], [ø]}, {[ɛ], [œ̃]}). So, the 1980 vowels are clustered into 12 vocalic classes (i.e., [a], [i], [e], [ɛ], [y], [ø], [u], [o], [ɔ], [ɑ̃], [ɛ̃], [ɔ̃]).

5.2. Vowel representation spaces

The tested representations of the vowels, from the most "elementary" to the most "sophisticated", are the following :

— Representation $W_{\text{Large Label}}$: the vowel is represented by the signal window centred on the temporal localization of the phonetic label (cf. §5.1.). This is a local representation corresponding to the "centre" of the sound realisation (i.e., generally the point of maximal intensity).

— Representation $W_{g(GTD,\ CSP)}$: the vowel is represented by the signal window centred on the centre of gravity of the interpolation function computed from the GTD technique using the CSP measure. This is a local representation deducted from the GTD model of the spectral evolution and which can be interpreted in terms of spectral characteristic of the given sound.

— Representation $S_{\phi(GTD,\ CSP)}$: the vowel is represented by the signal segment defined by the temporal support of the interpolation function computed from the GTD technique using the CSP measure. This is a global representation equally deducted from the GTD model and which reflects the segmental contribution of the given acoustic event.

5.3. Experiments on symbolic learning Learning and rule-based decision principle

For each vowel, the learning set is constituted from the representation of the first occurrence of the vowel [V] in the context $[C_nV]_{(n=1,..,16)}$, (i.e., 16 references per vowel). Complementary of the learning set, the test set is composed of the 1789 left vowels of the database.

Charade doesn't allow the learning of a signal continuum. Consequently, among the selected representations of vowels only the local representations can be tested for the learning. So, two rule systems are effectively generated by Charade from the representations $W_{\text{Large Label}}$ and $W_{g(GTD,\ CSP)}$. A third one is the simple concatenation of the previous rule systems.

The rule-based decision principle consists in taking a decision according to the number and the type of rule firings on the test-entity to identify. To this end, a counter is associated to each vocalic class. Each generated rule concludes on a terminal condition : the phonetic label of a vowel. At the identification step of a test-entity : for each rule of the generated rule system, a firing of rule on the test-entity increments the counter associated to the terminal condition of this rule. The highest counter, if strictly higher, involves an accuracy or a substitution, otherwise a rejection.

The results of recognition (cf. Table 1) are given for each selected rule system and for two representations of the test-entity : the first one (I) is local (i.e., $W_{\text{Large Label}}$ or $W_{g(GTD,\ CSP)}$) and the second one (II) is global (i.e., $S_{\phi(GTD,\ CSP)}$).

Charade Learning_$W_{\text{Large Label}}$	Accuracy Rate	Substitution Rate	Rejection Rate
I Test_$W_{\text{Large Label}}$	66,6 %	31,7 %	1,7 %
II Test_$S_{\phi(GTD,\ CSP)}$	65,9 %	33,6 %	0,5 %
Charade Learning_$W_{g(GTD,\ CSP)}$	Accuracy Rate	Substitution Rate	Rejection Rate
I Test_$W_{g(GTD,\ CSP)}$	64,4 %	34,1 %	1,5 %
II Test_$S_{\phi(GTD,\ CSP)}$	65,7 %	34 %	0,3 %
Charade Learning_$W_{\text{Large Label}}$ & Learning_$W_{g(GTD,\ CSP)}$	Accuracy Rate	Substitution Rate	Rejection Rate
I Test_$W_{\text{Large Label}}$	67,3 %	32,1 %	0,6 %
II Test_$S_{\phi(GTD,\ CSP)}$	66,9 %	33 %	0,1 %

Table 1. Results of experiments on symbolic learning

Identification on test-segment (II) is not shown to be relevant relatively to a local identification (I). The centres of gravity of the interpolation functions of GTD technique don't prove to be a strategic information for identification. The enhancement of rule system increases slightly the accuracy rate whatever the representation of test-entities is.

5.4. Experiments on statistical methods Measures and nearest-neighbor decision principle

The goal of these experiments is to compare the results of recognition obtained from symbolic learning with the results of usual statistical method. The method we choose is the nearest-neighbor in the meaning of LPC-based measure between the test to identify and the references of the learning set. Issued from the same auto-regressive model (i.e., a 16th order linear prediction over 25.6 ms windows), various LPC-based measures are selected for this comparison : euclidean metric on Log Area Ratio Coefficients [11], euclidean metric on Linear Prediction Cepstrum Coefficients [7], Log Likelihood Ratio measure [8] and Clustering of Spectral Peaks measure (cf. §2.).

For these experiments, the selected representation of vowels is $W_{\text{Large Label}}$. The learning and test sets are those of the previous experiments (cf. §5.3.). The results of recognition (cf. Table 2), split in two categories (i.e., accuracy and substitution) according to the nearest-neighbor decision, are given for each selected measure. For the similar experiment (i.e., Learning_$W_{\text{Large Label}}$ and Test_$W_{\text{Large Label}}$), the Charade's results are given again.

Nearest-Neighbor Decision	Accuracy Rate	Substitution Rate
Log Area Ratio Coefficients	55,1 %	44,9 %
Linear Prediction Cepstrum Coefficients	61,7 %	38,3 %
Log Likelihood Ratio	66 %	34 %
Clustering of Spectral Peaks	**70,1 %**	**29,9 %**
Charade Symbolic Learning	**66,6 %**	**31,7 %**

Table 2. Results of experiments on statistical methods

We note an advantage for CSP measure on the others LPC-based measures. The Charade's results are quite comparable to statistical results (i.e., results ranked between LLR results and CSP results). These results are very encouraging for further experimentations on symbolic learning.

VI. CONCLUSION

We have presented an approach to take into account the temporal evolution of the speech signal in a symbolic learning and rule-based recognition process. The experiments have shown that a priori strategic information, such as localization or segmentation obtained from GTD technique, isn't useful information for the vowels identification. Similar experiments on the selected statistical methods, using dynamic time warping on segment identification, have confirmed this results analysis.

The comparative study of Charade and usual statistical method is very encouraging for further research on symbolic process. Indeed, the transformation from numerical into symbolic is very problematic. The example description we proposed is particularly elementary. There is no doubt about the existence of a more relevant description. For instance, the enhancement of the original description (cf. §4.3.) increases about 4 % the accuracy rate. Efforts must be concentrated on the research of an improved description. To this end, the study of numerical-symbolic transfor-mations and the pursuit of the research on invariant cues seem to be essential. The success of such a research should show that symbolic learning is a concurrent approach to statistical methods. Other important advantages of this approach haven't been discussed in this paper. For instance, the recognition time cost without any common measure with the other methods and the capability to analyse the production rule system for knowledge acquisition.

REFERENCES

[1] B.S. Atal, "Efficient Coding of LPC Parameters by Temporal Decomposition", *IEEE ICASSP*, pp. 81-84, 1983.

[2] L.-J. Boë & L. Miclet, "Manuel d'étiquetage large. GRECO n°39 Communication Parlée", *Commission Etiquetage BDSON & EUROM*, 1988.

[3] L.-J. Boë, P. Perrier & G. Bailly, "The Geometric Vocal Tract Variables Controlled for Vowel Production : Proposal for Constraining Acoustic-to-Articulatory Inversion", *Journal of Phonetics*, n°20, pp. 27-37, 1992.

[4] M.-J. Caraty, "Contribution au décodage acoustico-phonétique. Etudes de distances inter-spectres et reconnaissance de cycles vocaliques", *Thèse de l'Université Paris 6*, 1987.

[5] M.-J. Caraty & C. Montacié, "Intégration de la décomposition temporelle généralisée dans un système d'apprentissage symbolique", *19èmes JEP*, pp. 387-392, 1992.

[6] J.-G. Ganascia, "Learning with Hilbert Cubes", *2nd European Working session on Machine Learning*, Sigma Press, pp. 158-171, 1986.

[7] A.H. Gray & J.D. Markel, "Distance Measures for Speech Processing", *IEEE TASSP*, vol. ASSP-24, pp. 380-391, 1976.

[8] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition", *IEEE TASSP*, vol. ASSP-23, pp. 67-72, 1975.

[9] C. Montacié, "Décodage acoustico-phonétique : apport de la décomposition temporelle généralisée et de transformations spectrales non-linéaires", *Thèse de l'ENST*, 1991.

[10] C. Montacié, M.-J. Caraty & X. Rodet, "Experiments in the Use of an Automatic Learning System for Acoustic-Phonetic Decoding", *ICSLP-90*, pp. 357-390, 1990.

[11] R. Viswanathan, J. Makhoul & W. Russel, "Towards Perceptually Consistent Measures of Spectral Distance", *IEEE ICASSP*, pp. 485-488, 1976.

[12] H. Ye, M.-J. Caraty, L.-J. Boë & D. Tuffelli, "Structural Phonetic Evaluation of Dissimilarities Functions Used in Speech Recognition", *Eurospeech*, pp. 404-407, 1989.

[13] H. Ye & D. Tuffelli, "Evaluation de distances en utilisant des sons synthétiques et la perception humaine",*15èmes JEP*, pp. 147-151, 1986.