TOWARDS AN AUTOMATIC LABELLING SYSTEM

Charles Barrera
Jacques-François Malet
Nadine Vigouroux

Laboratoire C.E.R.F.I.A.
UA 824-CNRS GRECO
118, Route de Narbonne
31062-TOULOUSE - F R A N C E

Jean Caelen
Geneviève Caelen-Haumont

ICP/INPG
LA 368-CNRS GRECO
46, Avenue Félix Viallet
38031-GRENOBLE - F R A N C E

ABSTRACT

We describe the environment required for a fine frequential labelling; i.e., the code and the operation systems resorted to. We also show how it is possible to devise a system that is capable of assisting an automatic labelling system.

I. INTRODUCTION

Certain problems, connected to Acoustic-Phonetic Decoding, call both for the elaboration of Acoustic Phonetic Data Bases (APDB) and —if only to constitute reference systems— for their labelling. Within the scope of the "Spoken Communication" G.R.E.C.O. (CNRS Coordinated Research Group), various mutually complementing approaches to labelling have been retained; e.g., broad, fine (both temporal and frequential) [1], [2], [3], normative phonetic transcription, etc.. In the present article, we describe the "environment" required for a fine frequential labelling; i.e., the code and operation systems resorted to. We also show how, thanks both to manual labelling and to an APDB, it is possible to devise a system that is capable of assisting an automatic fine frequential labelling. In order to do this, we use phonetic units to set up a correspondence between strips of both signal and spectrum, so that information items —that are useful to both learning and assessment procedures— can in due time be extracted.

II. THE ACOUSTIC MODULE

The first step, in fine frequential labelling, consists in achieving a spectral analysis of the vocal signal. The module of acoustic processing, we have on hand, is derived from a filter bank [4] : it yields spectrum in decibels on a 24-channel MEL scale. One spectral sample corresponds to a 128-dot window of analyzed signal; therefore, over a duration of 8 ms. at a 16 kHz sampling frequency. In practice, this is the one analysis we use, although we have on hand other methods —e.g., FFT, Cepstrum, LPC.

In order both to interpret and identify the signal, the expert makes use of several types of parameters:

-instantaneous values of: i)signal energy in dB (measured immediately after pre-stressing due to the ear-model), ii)formants, iii)spectral cues [5] and iv)fundamental frequency.

-temporal evolution both of the above parameters and of the Continuous/Discontinuous cue that measures the spectral derivative.

Once the parametrization system is specified, there remains to define the unit which the expert is going to work with. The unit we retained is the homogeneous infra-phonemic segment. At this level, processing is entirely automated.

The ' boundaries of the homogeneous-segment unit are determined through a segmenting function, computed on the basis of the overall variation of the acoustic and prosodic cues —using a modified version of delta coding [6]. A boundary is automatically set, whenever the segmenting function happens to exceed a certain threshold value —that is variable and decreasing with time— or whenever the unit exceeds 60 ms. in duration. The unit, thus obtained, is bound to remain smaller than the phoneme, whatever the value reached by the segmenting function.

III. LABELLING METHODOLOGY

Labelling consists in placing a set of codes, either directly onto the signal in the case of temporal labelling, or onto the spectrogram in the case of frequential labelling. True enough, the temporal domain is still favored by phoneticians: the raw signal is devoid of mathematical processing, that is an unfailing source of alterations. Nevertheless, we chose spectral reading:

-on the one hand, within a spectrum, phenomena such as nasal murmur or friction can both be detected,

-on the other hand, the automated definition of homogeneous segments renders the phonetician's task easier.

The codes to be placed can concern a whole segment [2] or, again, they can be used to spot events accurately [1]. Our approach is segmentwide: the expert places a set of labels —whose definition is underpinned by a phonetic model— at the boundaries of the "automatic" segments, defined above.

III.1 The System of Codes

III.1.1 Definition of Label Vector-Components

Acoustic, phonetic and syllabic properties are characterized through a label vector that is made up of several components, placed according to a previously set positional order. The system of codes consists of six different classes; five of which are set and only one allowed to vary:

-two classes —macro-class, C1, and phoneme code, C2— aim at phoneme characterization,

-two other classes, more closely related to the homogeneous segment, help in further specifying C1 and C2 above. The class dealing with acoustic phonetic modality, C6, is left to vary (i.e., several simultaneous descriptive adjectives are allowed). Class C4 consists of contextual attributes, all coined to give an account of co-articulation phenomena.

-a further class —acoustic phases, C3—sets segments in sequence within a given phoneme realization,

-a final class, C5, supplies information at the syllable level, depending upon the position occupied by a syllable within larger conceptual entities; e.g., word, wordgroup, phrase.

Let us now take up each such class, in the order the expert follows while labelling:

C1: Macro-Class

We recognize ten distinct macro-classes:
Vowel (V), Nasal Vowel (M), Semi-vowel (W), Liquid (L) covering /l/ and /r/, Nasal Consonant (N), Voiceless Occlusive (Q), Voiced Occlusive (O), Voiceless Fricative (S), Voiced Fricative (Z) and Silence (P).

C2: Phoneme Codes

These codes take after the International Phonetic Alphabet (IPA), which are not available on computer keyboards; e.g., /ã/ coded **AN**, /ɔ/ coded **O**, etc.

Class C1 is redundant, with respect to the phoneme code C2; indeed, it is automatically generated by the system, whenever the expert identifies a phoneme.

C3: Acoustic Phase

This component describes the temporal unfolding of homogeneous segments within a given phoneme: Onset or Establishment phase (E), Sustained or Steady phase (T) and Coda or Phaseout (Q). These three phases are applied systematically to both consonants and vowels.

C4: Contextual Attributes

These describe how contextual events concur with or, sometimes even, prevail over the expected realization of a phoneme; staking phenomena that often more properly pertain to phonology. The cases most often encountered are: **A** for Approximating (e.g., final /ə/), **R** for Substituting, **I** for Insert, **F** for Merging.

C5: Syllable Delimitors

These codes are defined over non-complex phrases that are considered in a twofold manner:

-along the axis of structural complexity: Phrase (P), Groupword (G), Syllable (I, S, F, M),

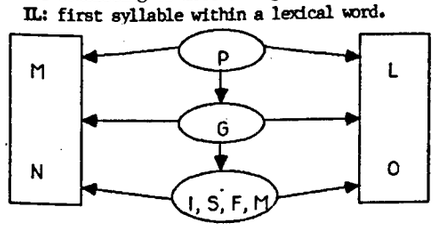
-along the lexical axis: L for Lexical word, O for Tool or Grammatical word.

An additional distinction is made between mono- (M, N) and pluri-syllabic (L, O) words. Examples:

PL: first syllable within a phrase that begins with a lexical word consisting of more than just one syllable,

GM: first syllable within a wordgroup that begins with a lexical word consisting of more than just one syllable,

IL: first syllable within a lexical word.



C6: Modality of Realization

Acoustic or articulatory modality specifies some of the implicit features pertaining to a macro-class (loss and/or addition and/or alteration of acoustic features). Since the field of modality is open to variability, it is possible to choose among a number of descriptive adjectives: Oral (O), Vocalic (V), Glottal (G), Nasal (N), Consonantal (C), Unvoiced (S), Semi- (2), Closure (K), Burst (X), Fricative (F), Palatalized (Y), Affricate (Z), Aspirate (H), Noisy (B). This list is by no means exhaustive, and can be updated should new acoustic features become pertinent.

Defining and placing these label-components appeared, at first, rather to be a matter of interpretation than one of description. However, little by little, there began to emerge a number of steady conventions, likely to sustain a more constant phonetic interpretation of acoustic facts; thus resulting in a type of labelling that is more descriptive than interpretative. Still, classes C1, C2 and C3 can be considered as belonging rather to the descriptive type, whereas C4 and C5 are definitely more subjective and are, therefore, a matter of interpretation. More specifically, Acoustic Phase (C3) is:

-at times, descriptive, and this is the case both of discontinuous vowels and of discontinuous consonants, both fairly easily opened to segmenting rules,

-at other times interpretative, and this is the case —delicate, if anything— of semi-vowels, over which the notion of phase applies with truly extreme difficulties.

III.1.2 Verifying the Labels

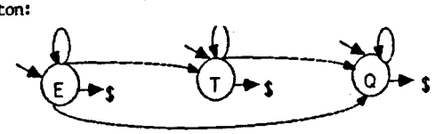
Errors, in implementing both the syntax and the semantics of labelling, can intervene in the course of manual labelling. Therefore, it is necessary to check the manually applied labels, at least for proper syntax.

The procedure is run in three steps:

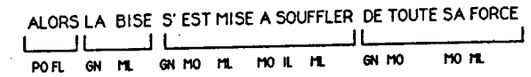
-Since the label vector is a pre-defined structure of components both belonging to a finite set of values and obeying to a strict positional order, it is procedurally possible to check that the value, specifically assigned to a component, does belong to the appropriate set of definition of such values. Thus typing mistakes, which necessarily cause improper labels to be entered, can be automatically detected and then removed.

-Within the temporal sequence, while shifting from one label to the next, choice of value is not arbitrary; indeed, it is subjected to sets of rules respectively applying to the different types of components. Successive values are not drawn, from one such set, in a random order. Thus, for example, when labelling for acoustic phase (although, in practice, labelling order is accross all class-defining sets to formulate one label at a time), sequences such as [E Q T] or [E T Q E], ..., over one and the same phoneme, are strictly prohibited.

The process rule can be symbolized through the following automaton:



Likewise, sequencing syllable codes within phrase structure can be managed by a similar automaton. Examples of syllable coding from the corpus of the G.R.E.C.O.'s Database of sounds [7]:



-The third verification step bears upon mutual compatibility among the components making up one and the same label and, more particularly, on combinations involving C1-C2-C3. Unauthorized combinations are those involving either a redundant modality (Ex.: VO; i.e.,vowel V-Oral modality) or a contradictory one (Ex.: by definition, macro-class "Q" excludes Oral modality "O").

III.2 The Operating System

III.2.1 The Spectrogram Editor

Labels placed by an expert are machine-acquired thanks to a spectrogram editor. This software makes it possible to listen to the signal, to view it, to display the corresponding spectrogram, as well as the various cue curves —viewable with or without zoom, to watch both formant values and fundamental frequency, ...

Such a system offers two major advantages:

a)Default labelling is necessarily infra-phonemic, since the labelling agent must set his/her/its view of reality in correspondence with the automatically determined segments; that is, in the case of fine labelling. The system is capable, as well, restrictively to deal with only phoneme designation; labelling over classes C1 and C2 only. Thus it is possible to label broadly (Likewise, it would be possible to concentrate on supra-phonemic units; e.g., diphones).

b)The editor can handle any system of codes. Two other systems are, at present, being used within the scope of specific studies (foreign languages [8] and phrase complexity [9]). A user has only to specify the structure of the wanted label vector, the set of codes to be used by each type of component and, of course, their syntax within a given datastructure.

III.2.2 Verification Procedures

Once labelling is over, a verification procedure is initiated on labels. Procedures, defined on the basis of concepts mentioned earlier (See III.1.1 supra), supply an opportunity for a correction that is interactive with the user. Such a module ensures both a quality- and a reliability-control of the labels produced for the database.

III.2.3 Recapitulating Example

In order to label, the expert has on hand:

-signal that can be both viewed and listened to,

-information items, displayed as spectrograms and curves. Description of these items: (cf. Figure 1)

From left to right, we can see:

-N, the spectrum sample number,

-W, the signal's energy in dB, with an evolution curve vs. time,

-homogeneous segments —whose boundaries are materialized by a number of "<"— are secured through automated segmentation.

-the 8 ms. skeleton spectrum,

-the cues, mentioned supra, displayed as histograms,

-the segmenting marks, properly speaking, that the expert places in correspondence with homogeneous segment (A very high proportion of automatically delivered homogeneous segments get one such mark).

Various zoom-pictures (cues, signal,...) are available, by request, on graphic screen.

IV. THE ACOUSTIC-PHONETIC DATABASE AND LABELLING

We now look at how labelling is closely knit in the elaboration of an APDB.

IV.1 Setting up an APDB

IV.1.1 Information Retained

In order to meet the goals, entailed in setting up a system that delivers automated fine frequential labelling, the necessary APDB must, in the course of manual labelling (system priming), assemble all the required information; namely, information encompassing all processing phases, from the physics of vocal signal to sophisticated linguistic notions.

Two kinds of information, however, should be distinguished:

-Quantitative data: signal samples, spectral samples, prosodic parameters (cf. § II above) and infra-phonemic segment boundaries.

-Qualitative data: labels corresponding to linguistic conceptual events that the expert detects in the course of manual labelling.

All such information is woven into the APDB, thanks to a management system [10].

IV.1.2 Relations

The management system aims at tying together the various types of information, described above.

Through a new datastructure, various kinds of data become associated. For example: signal block number, spectrum sample number, label vectors as placed by expert... Thus, thanks to semantic links, items of symbolic information (phonetic concepts) become associated with items of acoustic quantitative information (spectrum, signal). In fact, this linking correspondence is one of the most crucial problems facing phonetic decoding. At least, this is an important assumption that the database scheme we propose, attempts to meet. Moreover, since users need to retrieve phonetic concepts from any context, it becomes useful to weave relations between the context that is being examined and other previous or subsequent contexts.

Thus, for example, given a phoneme it is possible to find out:

-its realized occurrence among blocks in the signal file,

-its occurrence within the centi-seconds of the spectrum file,

-its phonemic context, both prior and posterior,

-its next realization within the same file.

The latter two types of relations are systematically created for each label vector component. This retrieval scheme, and the set of links it entails, lies at the very base of any APDB consultation.

IV.2 Consulting the APDB

In order to learn an automated labelling system, retrieval of contents from various types of file is imperative (e.g., physical sounds, spectrum, labels issued from expertise, ...). Therefore, to the effect of facilitating new applications —i.e., learning procedures— we designed a retrieval sytem to reach all information elaborated from the vocal signal. This type of consultation is made possible, thanks to the semantic links that allow access both to units pre-defined through labelling and to relations between various datatypes. Thus, it becomes realistic to set up references such as:

-mean value of energy parameter over realizations of phoneme /i/ within a given corpus, uttered by a given speaker,

-mean value of fricative formant over realizations of F-class vowels (displaying the fricative modality),

-mean value of energy parameter over all T phases (sustained portion) of all occlusives within corpus,

-etc.

IV.3 Database Contents

At present, we have available an initial acoustic-phonetic database, labelled for 10 speakers (7 males, 3 females). The corpuses used are:

-connected digits and logatoms CVCVCV, both for C.N.E.T. Agreement [11],

-continuous speech: "La Bise et le soleil", for G.R.E.C.O., for a total of 13000 phonemes, labelled.

V. TOWARDS AN AUTOMATED LABELLING

The goal we set for ourselves is to help the phonetician's expert work. We mean either to automatize certain tasks or to further the degree of automatization, already achieved within the pre-segmentation module that yields homogeneous segments.

As an initial step, we limit our scope to the identification of both phoneme (C2) and modality (C6) components of a label vector. In the way of system input, we already have a normative phonetic transcription and a set of quantitative items of information (spectrum, signal) concerning any sequence we wish to label. From this transcription, we contemplate both introducing automated alignment procedures [12], [13], [14], [15] and comparing these with procedures that segment for events [16].

By automatically placing boundaries, such procedures should make it possible to delimit phonemes. Meanwhile, for the purpose of fine labelling, it is equally advantageous to add procedures for extracting acoustic and phonetic features (C6). Specifications for this phase must include:

-not only a strategy of expert labelling [17], [18], [19],

-but also learning results delivered by statistical modules, when these are run on a base of already labelled data.

For the time being, the system is devised both to validate labels, and as a tool serving expertise. Next, we mean to formalize our results, with a view to elaborating an automated interactive labelling sytem.

ACKNOWLEDGMENT

Our warm thanks to Dany Laur who joined us on the lengthy spectrogram-reading expert's task.

VI. BIBLIOGRAPHIC REFERENCES

[1] C. Abry, C. Benoît, L.J. Boé, R. Sock, "Un Choix d'Evénements pour l'Organisation Temporelle du Signal de Parole", **XIV JEP, GALF-CNRS**, Paris, Juin 1985, pp. 133-137.

[2] D. Autesserre, M. Rossi, "Propositions pour une Segmentation et un Etiquetage Hiérarchisé. Application à la Base de Données Acoustiques du GRECO- CP", **XIV JEP, GALF-CNRS**, Paris, Juin 1985, pp. 147-151.

[3] C. Abry, D. Autesserre, C. Barrera, C. Benoît, L.J. Boé, J. Caelen, G. Caelen-Haumont, M. Rossi, R. Sock, N. Vigouroux, "Propositions pour la Segmentation et l'Etiquetage d'une Base de Données des Sons du Français", **XIV JEP, GALF-CNRS**, Paris, Juin 1985, pp. 156-163.

[4] J. Caelen, "Space/Time Data-Information in ARIAL Project Ear Model", **Speech Communication**, Vol 4, Aug. 1985, pp. 163-179.

[5] J. Caelen, G. Caelen-Haumont , "Indices et Propriétés dans le Projet ARIAL II", **Actes du Séminaire Encodage et Décodage Phonétiques, GALF-CNRS**, Toulouse, 1981, pp. 129-143.

[6] N. Vigouroux, J. Caelen, "Segmentation Phonétique et Organisation d'une Base de Données Acoustiques et Phonétiques", **XIV JEP, GALF-CNRS**, Paris, Juin 1985, pp. 152-155.

[7] R. Descout, J.F. Sérignat, O. Cervantes, R. Carré, "Une Base de Données des Sons du Français", **12-th ICA**, July1986.

[8] J.F. Malet, "Une Méthode Acoustico-Phonétique pour l'Enseignement Automatique de Langues Etrangères", **XV JEP, GALF-CNRS**, Aix-en Provence, Mai 1986.

[9] G. Caelen-Haumont, "Grammatical Components and Macro-Prosody: Quantitative Analysis Toward Statistical Correlations", **12-th ICA** , Toronto, July 1986.

[10] J. Caelen, N. Vigouroux, "An acquisition and Research System for an Evolving Nucleus of Acoustico-Phonetic Knowledge", **IAPR, 8-th ICPA, AFCET**, Paris, 28-31 Octobre 1986.

[11] **Convention CNET N° 86 7B 020**, "Réalisation et Exploitation d'un logiciel de validation d'indices Acoustiques pour la Reconnaissance de la Parole Multi-Locuteur".

[12] J.S. Bridle, R.M. Chamberlain, "Automatic Labelling of Speech Using Synthesis-By-Rule and Non-Linear Time-Alignment", **11-th ICA**, Toulouse 1983, pp. 187-189.

[13] A. Andreewsky, M. Desi, C. Flur, F. Poirier, "Une méthode de Mise en Correspondance d'une Chaîne Phonétique et de sa Forme Acoustique", **11-th ICA**, Toulouse 1983.

[14] P. Collins, S. Barber, "Fine Phonetic Labelling Methodology for Speech Recognition Research", **Proceedings IEEE-ICASSP**, Tokyo 1986, pp. 2779-2782.

[15] H.C. Leung, V.W. Zue, "A procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech", **Proceedings IEEE-ICASSP**, San Diego, 1984, pp. 2-9.

[16] G. Pérennou, M. de Calmes, "Segmentation en événements phonétiques et unités syllabiques", **XIV JEP, GALF-CNRS**, Paris, Juin 1985, pp. 142-146.

[17] F. Lonchamp, "Reading Spectrograms: The Viewfrom the Expert", in **Fundamentals in Computer Understanding: Speech and Vision**, ed. J.P. Haton, Cambridge University Press, 1987, pp.181-206.

[18] N. Carbonnel, D. Fohr, J.P. Haton, F. Lonchamp, J.M.Pierrel, "An Expert-System for the Automatic Reading of French Spectrograms", **Proceedings IEEE-ICASSP**, San Diego 1984, pp. 42-8.

[19] P.E. Stern, M. Eskenazi, D. Memmi, "An Expert System for Speech Spectrogram Reading", **Proceedings IEEE.-ICASSP**, Tokyo,1986.

FIGURE 1