The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

AUTOMATIC LABELING OF LARGE PROSODIC DATABASES : TOOLS, METHODOLOGY AND LINKS WITH A TEXT-TO-SPEECH SYSTEM

Gérard Bailly, Thierry Barbe & Haï-Dong Wang

Institut de la Communication Parlée, Unité Associée au CNRS N° 368
INPG/ENSERG, 46, av. Felix Viallet. 38031 Grenoble CEDEX, FRANCE

ABSTRACT

This article presents an unified methodology to segment and label acoustic databases. The methodology is entirely based on a phonetic model : the temporal decomposition (TD) model. In this model phonemes are seen as emergence functions (EF) which overlap in time. The segmentation and the determination of the prosodic contour of an acoustic continuum is intimately linked with the detection of the EFs. As the EFs are automatically determined the coherence of the prosodic structure of utterances across the entire corpus is ensured and thus statistical methods can be applied to study the links between formal analysis of the text and prosodic structure of the message. Since the same methodology may be applied to the segmentation of phonetic units, synthesis by concatenative units may be performed : prosodic events detected in the prosodic database and in the phonetic units are entirely compatible. The tools presented below are speaker-independent and cover the entire analysis to synthesis process.

1. INTRODUCTION

The creation of improved synthetic models of prosody implies the extensive collection of data organized into a large database. The content of these databases is studied carefully in order to focus on the prosodic phenomenon one want to study (relation with syntax, semantics, speech rate...). At that stage we face multiple problems: first one is the time-consuming task of labeling the database, the second one is the maintenance of the coherence of this database. Manual segmentation (and so determination of phonemes durations) has a high variability and is often not reproducible. Furthermore we need an objective criterion related to phonetic events to be able to use the results into a synthesis system in a coherent way.

Once the synthetic model has been build using statistical measurements [Carlson & Granström, 1986] or phonological models [Bailly, 1989], we face the problem of evaluating this model. It should be tested against original data of the database and also in a prediction task of new data. This test have to be as much separated from segmental problems as possible in order to focus only on prosodic quality. We present here a coherent methodology and the associated tools to perform high-quality construction and testing of large prosodic databases. Automatic prosodic labeling is performed using speaker-independent phonetic alignment based on temporal decomposition of speech (referenced as TD [Bailly & al, 1989]) and a robust pitch tracker. Test of synthetic prosody can be achieved thanks a PSOLA synthesizer controlled by high quality period marker and rhythm controller based on TD.This PSOLA synthesizer may prosodically "deform" the original data as well as concatenation of synthesis units coming from a segment dictionary and thus originate a text-to-speech synthesizer.

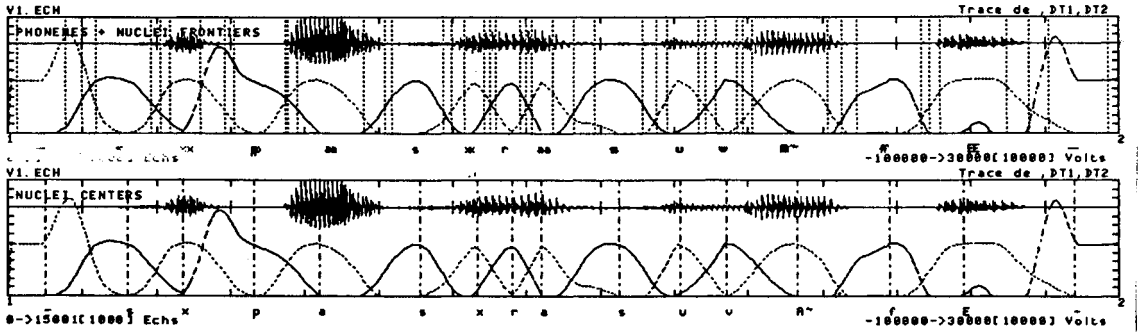

Fig.1. Automatic segmentation of the sentence: "ce pas sera souvent fait" (this step will be often done). From the bottom to the top are presented : i) the signal with the NCs, ii) the signal with the OTs, BSs and ESs. The optimal EFs obtained by DTW are superposed.

2. METHODOLOGY FOR THE CONSTITUTION OF PROSODIC DATABASE

The information stored into the database has to be rich enough to enable high-quality resynthesis of the stylized versions of the prosodic contours while these contours must be defined with a restricted number of values to enable their association with phonological descriptions of the involved message.

Our methodology of constitution of prosodical databases is based on an underlying phonetic model : each phoneme is associated with an EF.These EFs vary between 0 and 1. They are coarsely bell-shaped and are the coordinates of the current spectral shape on the plan defined by the two

surrounding targets considered as independent. Due to the well-known coarticulation phenomenon [Bailly & al, 1989], these EFs overlap in time with the adjacent ones. These overlaps could extend more than one phoneme but we have chosen to limit the model to only two interleaving functions. These EFs are highly correlated to spectral masses movements.

All measurements stored into the database are related to time events related to these EFs. The time events we selected are (cf.Fig.1):

- the onset time (referenced as OT) defined as the time where EFs of adjacent phonemes are equal. This event is used for phoneme duration determination.
- the onset of the stationary part (referenced as BS) defined as the time of the point at 0.8 of the onset edge of the current phoneme.
- the offset of the stationary part (referenced as ES) defined as the time of the point at 0.8 of the offset edge of the current phoneme.
- the nucleus center (referenced as NC) defined as the time of the center of gravity of the stationary part.

Following recommendations defined in [Emerard & Benoit, 1988], we have chosen to store for each phoneme six values :

- duration (defined as the time between the OT of the current phoneme and the next OT,
- nucleus duration (defined as ES-BS) and 3 F0 values: at BS,NC and ES,
- nucleus energy : average of the signal energy on the interval [BS,ES].

Fig.2. gives an example of the desired output processed by the automatic procedure presented below for a test sentence. This output is obtained using the following methodology which can be divided into 4 main steps:

- automatic nucleus determination and labeling involving EFs calculation,
- time events determination and F0 calculation using robust pitch tracker [Barbe & Bailly, 1990],
- storage of the prosodic values at the corresponding time events.

Nom	Duree	Noyau	F0	F0	F0	EnrdB
-	100	66	0	0	0	34
s	105	74	0	0	0	49
x	87	41	189	167	151	74
p	83	28	0	0	0	36
a	134	90	218	230	245	81
s	89	69	0	0	0	54
x	53	27	175	161	157	74
r	49	32	148	143	144	71
a	65	29	146	146	154	73
s	111	65	0	0	0	49
u	67	43	195	197	197	75
v	51	28	195	192	186	63
A~	132	101	189	182	165	73
f	119	87	0	0	0	40
E	133	90	167	132	106	70
-	122	81	0	0	0	34

Fig.2. : Prosodic file for the sentence of Fig.1. On each line is presented the phoneme, the phoneme and the nucleus duration, the 3 F0 points and the nucleus energy.

3. TOOLS FOR DATABASE LABELING

The tools developed for that application aim robustness and speaker-independence. No training phase is then required for a new speaker while no speaker-specific thresholds have to be specified.

3.1. Phonetic labeling : marking phonemes nuclei

The phonetic labeling is performed in three stages : a) pre-segmentation b) alignment and c) nucleus centers adjustments. The evaluation has been done using a reference segmentation done by hand using usual spectral criteria. A nucleus was considered good if it lies alone inside a segment labeled with the same phonetic label.

3.1.1. Pre-segmentation

The pre-segmentation is largely inspired from Philips work [van Hemert, 1987]. Instead of selecting transient candidates we use this center-of-gravity method to output nucleus candidates. A rate of 86% of good nucleus determination was reached (only one event inside one segment) with 8% omission and a rate of 2.3 events per insertion.

3.1.2. *Alignment*

This stage has in charge to align the phonetic string supposed correct with the event candidates issued by the pre-segmenter. We use the usual dynamic time warping (referenced as DTW)

procedure with appropriate local distances and transition costs. Backtracking will then select, delete or insert events and associate them with labels. The global cost encounters minimization of the reconstruction errors with the hypothesized EFs and maximization of local phonetic probabilities calculated with histograms on energy and zero-crossings.

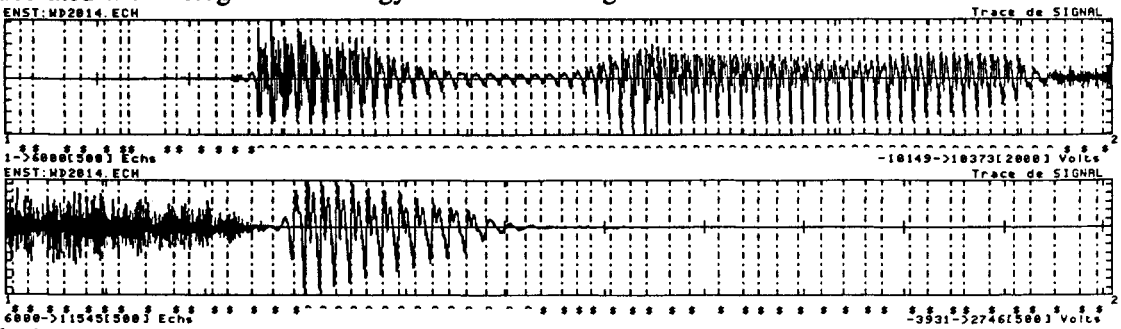

Fig.3. Pitch markers generation for the logatom /azaeʃu/. Voiced periods are pointed by the ^ sign and unvoiced ones by the $ sign.

3.1.3. Nucleus centers adjustment

Based on the optimal EFs determined by the precedent procedure new nucleus position are computed. These new positions are the centers of gravities of the corresponding EFs.

3.2. Fundamental frequency estimation and pitch markers generation

F0 estimation and pitch markers generation is achieved using a robust pitch tracker working indifferently on the speech wave or the signal issued by a glottograph. This robust pitch tracker operates left-to-right on the speech signal in only one pass. It detects beginnings and endings of voiced parts using pattern recognitions techniques on the energy and zero-crossing contours. At the beginning of the voiced parts F0 and pitch markers anchor points are calculated using AMDF transformation and DTW [Barbe & Bailly, 1990]. On a 40 ms interval after the voiced onset AMDF transformation at a frame rate of 200Hz outputs 8 rows of F0 candidates. The right F0 and pitch markers contours are selected using DTW. The algorithm proceeds until the end of the voiced part using a data reduction technique. On unvoiced segments, pitch markers are generated randomly following distribution of the voiced periods. All pitch markers coincide with zero-crossings. Fig.3. presents the result of the algorithm on the logatom /azaeʃu/: randomly generated pitch markers are pointed by the $ sign.

4. METHODOLOGY FOR THE EXPLOITATION OF THE DATABASE

The methodology developed earlier produces a full labeled database. This information is the basis of a synthetic prosodic model. We are not concerned in this article with the different methodologies available to describe the underlying prosodic facts. But the description of this methodology to produce the database would not be complete if we hadn't proposed something to test any kind of synthetic model one can deduce from these data.

The synthetic model will produce a file with the same data structure as the database. The tools we propose in the following aim to produce a high-quality synthetic signal whose prosodic patterns are given by that file but whose segmental patterns are minimally distorted compared to a natural continuum, especially the continuum of a database sentence.

4.1. PSOLA: General synthesis-by-analysis technique

We have developed a synthesis-by-analysis technique based on the PSOLA paradigm [Hamon & al, 1989]: adaptation of a natural speech to different prosodic parameters can be obtained with minimal spectral distortion using simple OLA (overlap-and-add) methods. Duration control is obtained by replication of synthetic periods of the signal. These synthetic periods of the signal are obtained from corresponding original periods by simple displacement of elementary speech signals centered on pitch markers i.e. middle of the closed phase of the period. These elementary speech signals consist of the two periods surrounding the current pitch mark weighted by windows with good spectral properties (usually Hanning or Blackman-Harris windows).

4.2. Control of the synthetic speech rhythm

A uniform duration change through the entire phoneme is dangerous : transient parts durations are governed mainly by articulatory constraints (precision of the articulatory target, distance from the preceding one...). Stationary parts seem to bear the most part of the prosodic control. Moreover stressed vowels have often shorter transient parts than unstressed vowels. We choose to correlate cadence control distribution with the EFs so that maximum period insertions or deletions occur during the nucleus.

4.3. Control of the synthetic speech energy contour

The synthetic energy contour is adjusted in the following way : a linear interpolation function is applied to the energy of each period between two adjacent NCs such that the synthetic energies of the periods corresponding to these two NCs equal to the energy requirements given in the prosodic file.

5. LINKING THIS METHODOLOGY WITH THE DEVELOPMENT OF A TEXT-TO-SPEECH SYSTEM BASED ON CONCATENATIVE SYNTHESIS

High-quality speech synthesis may be obtained by following a similar methodology for, i) the automatic segmentation, labeling and pitch marking of logatoms or carrier sentences from whom segments of speech may be extracted (diphones, syllables...), ii) the prosodic control of the original continuum made up by concatenation of these minimal units. Automatic segmentation of synthesis units has been proven [van Hemert, 1987] to be more efficient than manual segmentation. Thus phonetic units may be extracted by selecting appropriate acoustic segments between two NC centers. The segment dictionary contains for each segment :

- the acoustic signal,
- the NC,OT,BS and ES times for each phoneme,the pitch markers and the energy of each period.

Synthesis is then comparable to the prosodic control of an original utterance which is build up by concatenation of these segments.

6. CONCATENATIVE SYNTHESIS FOR FRENCH

The methodology described above has been applied for French using a dictionary of "polysounds" coming from ENST laboratory in Paris. Polysounds are defined as the portions of signal going from the NC of a stable segment to the NC of the next one : all phonemes except semi-vowels (/ɥ/,/w/,/j/) and liquids (/l/, /ʀ/,/r/) are considered as stable. A complete inventory of the French polysounds needs around 3200 logatoms to be recorded. The maximum length of polysounds foı French is 5 for segments like /arlɥi/ or /ylrwa/.

The complete segmentation and labeling of the database was accomplished within 30 hours of CPU-time on a microvax II computer. A manual control of half of the polysounds give 95 % of correct labeling (deviation of manual NC centers inferior to 20 ms). Most of the errors are explained by the incorrect pronunciation of logatoms by the speaker and by the non-systematic adjunction of mute /ə/ at the end of logatom-final consonants. The intelligibility of the synthetic speech has not been yet systematically evaluated but the quality of the synthesis by polysounds driven by the prosodic description of original sentences uttered by the same speaker is comparable to the original sentences quality.

7. CONCLUSION

A methodology for an unified prosodic analysis and prosodic deformation of any acoustic segment has been presented : it gives a coherent prosodic description of any speech continuum with associate tools. The analysis and synthesis process is speaker and language independent. The size of the acoustic units are not constrained and the synthesis process may mix up analysis-synthesis of sentences or words with text-to-speech synthesis by concatenative units using the same prosodic control principles and thus give a comparable speech quality output.

We are currently investigating new synthesis techniques such as parametric PSOLA. A pitch-synchronous closed-phase LPC analysis followed by formants and bandwidths extraction and a PSOLA technique applied onto appropriate parts of the residual offers nearly the same speech quality while preserving an intelligent control on the spectral of the envelope.

REFERENCES

Bailly G., Marteau P.F. & Abry C. (1989), "A new algorithm for temopral decomposition of speech. Apllication to a numerical model of coarticulation", IEEE Conf. on ASSP, 508-511.

Bailly G. (1989), "Integration of rhythmic and syntactic constraints in a model of generation of French prosody", Speech Communication, 8, 137-146.

Barbe T. & Bailly G. (1990), "Evaluation d'un détecteur de fréquence fondamentale d'un signal microphonique par comparaison avec la mesure effectuée sur le signal laryngographique",SFA, 18ème JEP, 165-169.

Carlson R. & Granström B. (1986), "A search for durational rules in a real-speech database", Phonetica, 43, 140-154.

Emerard F. & Benoit C. (1988) "Base de données prosodiques pour la synthèse de parole", J. d'Acoust., 1, 303-307.

Hamon C., Moulines E. & Charpentier F. (1989), "A diphone synthesis system based on time-domain prosodic modifications of speech", IEEE Conf. on ASSP, 238-241.

van Hemert J.P. (1987), "Automatic segmentation of speech into diphones", Philips Tech. Review, 43,9, 233-242.

Wang H.D., Bailly G. & Tuffelli D. (1990), "Automatic segmentation and alignment of continuous speech based on the temporal decomposition model", JASA, 87,S105.