Evaluating the adequacy of synthetic prosody in signaling syntactic boundaries: methodology and first results.

Yann Morlec, Albert Rilliard, Gérard Bailly & Véronique Aubergé
Institut de la Communication Parlée
46, av. Félix Viallet 38031 Grenoble CEDEX FRANCE
(morlec,rilliard,bailly,auberge)@icp.inpg.fr

Abstract

This paper presents a methodology for evaluating the performance of synthetic prosody. The methodology focusses on the functional equivalence of natural and synthetic prosodic contours in terms of both distinctiveness and coherence. A perceptual experiment was designed to test the performance in terms of distinctiveness for a simple phrasing task. Results demonstrate the importance of a phonological representation that dynamically encodes fine details of the syntactic structure.

Introduction

Prosody shares at least two major functions with other linguistic processes (lexicon, morpho-syntax…): segmentation and hierarchical structuration. Depending on the type of speaking style, prosody is more or less related to syntax. In a framework of isolated utterances read in French, both structures are close, and the generation of an adequate mapping between these two structures is therefore a fundamental aspect of prosody modelling. The prosodic module of a text-to-speech system has thus to convert a set of symbolic descriptors (SD) caracterising syntactic boundaries into time-varying prosodic contours (PC). Rule-based or data-driven approaches can be used to approximate the real mapping function associating SD and PC by a predicted one. The major issue of evaluation is to quantify the *functional equivalence* between these two mappings.

The evaluation method

Minimising the distance between observed contours and predicted ones on a test set does not guarantee that predictions carry adequately the linguistic task defined by SD. Listeners are indeed not equally sensitive to all aspects of the prosodic continuum: different prosodic contours may indeed be "perceptually equivalent" ('t Hart, Collier & Cohen, 1990) or moreover "functionally equivalent" i.e. carrying the same SD via different contours - or variants.

Judging the functional equivalence of two contours is however a difficult task as prosody contributes to the structuration of the message in cooperation with the a priori knowledge of the listener about the message content, the language being used and the situation. The classical way of quantifying the *intrinsic* gain of intelligibility offered by "pure" prosody consists of degrading linguistic, acoustic or phonetic dimensions of speech while

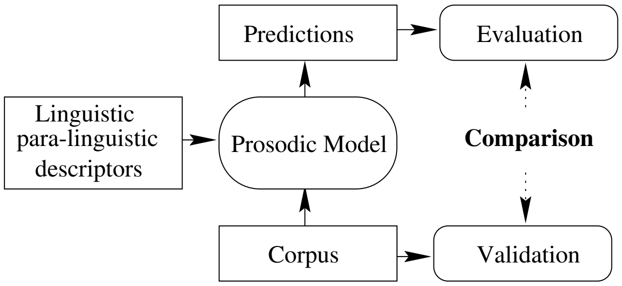


Figure 1: Methodology for evaluating prosody: a corpus is designed to reveal how the segmentation function assumed by prosody is encoded by a given speaker using given lexical items. This corpus is validated and its prosodic parametrisation is used to train a prosodic model. The model converts the symbolic descriptors into predicted prosodic contours. The output of the model is then evaluated according to the same procedure as used for natural stimuli.

preserving tested prosodic information. The EAGLES report (Gibbon, Moore & Winski, 1997) mentions three main methods: the Semantically Unpredictable Sentences series of tests (Benoît, Grice & Hazan, 1996), the degradation of segmental quality test (Carlson, Granström & Klatt, 1979) and the delexicalisation test (neutralisation of the phonetic content) e.g. glottogram (Lhote, Filleau & Grange, 1975), reiterant speech with a canonical syllable (Liberman & Streeter, 1978) or with allophones of different phonemes (Strom & Widera, 1996).

Procedures such as delexicalisation clearly minimise top-down contribution in the perception of phrasing. Listeners are however involved in very unnatural identification tasks that certainly increase their cognitive load (Pisoni, 1997). On the contrary we propose here a functional test that preserves linguistic, acoustic and phonetic dimensions of original stimuli except the controlled prosodic parameters. Evaluation of synthetic prosody is carried out in a relative way by comparing its performance with natural prosody (see validation-evaluation procedures shown in figure 1). The comparison is done in terms of adequation. In our approach, adequate prosody should structure the message the way speakers do in the target situation, maintaining distinctiveness only when necessary as suggested by Lindblom and colleagues (Lindblom & Lindgren, 1985).

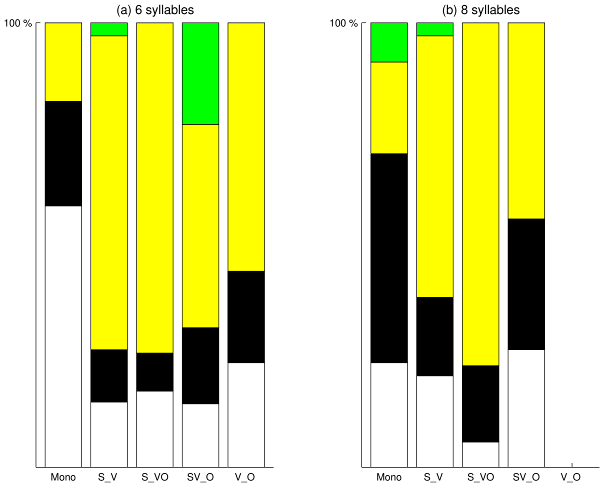


Figure 2: Direct comparison of natural vs. predicted "true" prosody. Average for 6 (left) and 8 (right) syllable utterances. Preference for (Nat) is shown in white and (Pred) in black, both accepted in light gray and both rejected in dark gray.

Here we apply this methodology to the phrasing component of our current superpositional model of intonation (Morlec, 1997).

Methodology

The system that generates prosody is seen as a system that converts linguistic structure into prosodic parameters. The test must enable the performance of the predicted mapping for the model user to be evaluated and must provide an internal diagnostic for the model builder.

The performance of the model is thus evaluated by a functional test that explores and compares the sensitivity of both natural and synthetic prosody to perturbations in linguistic specification - here syntactic parsing: replacing SD_i by SD_j. The perturbations are simulated by choosing pairs of sentences with similar phonotactic and phonetic dimensions but fulfilling two different linguistic specifications. Natural and synthetic prosody corresponding to each specification is then re-applied to the original sentence and also to the second one. Acceptance/rejection of these pairs of stimuli - coherent/incoherent associations between linguistic specification and prosodic phrasing - will help us to diagnostic (a) which level of granularity (size of SD set) for the linguistic specification is optimal, (b) what details of the linguistic description and other dimensions are missing in the input parameters and (c) which distinctions the prosodic model fails to capture.

Experiment

Principles

We limited this first application to the prosodic phrasing of declarative read French utterances with 4 simple syntactic structures: a single polysyllabic word, subject-verb, subject-verb-object and verb-object. All sentences exept the one containing the polysyllabic word have a syntactic boundary placed after the third syllable. This boundary thus separates different constituents.

Corpus

The test corpus consists of 5 sentences of 6 syllables and 4 sentences of 8 syllables (see Table 1). Note that the structure verb-object is not represented for 8 syllable sentences. Natural and synthetic prosody is not applied to original utterances in order to avoid an obvious preference for original associations. We used a set of utterances pronounced by the reference speaker with monotonous intonation - isochronous and with a flat melody ($94.82 \pm 0.54Hz$). Stylised natural and synthetic stimuli with specific melodic and rhythmic contours are then obtained by using a high-quality TD-PSOLA analysis-synthesis of these "monotonous" utterances.

Protocol

In a test ABBA, 18 subjects heard 146 pairs of utterances and had to judge which item was the most adequate for a neutral reading of the sentence written on the screen. If they were not able to choose clearly, they were allowed to answer that both - or none - were adequate. Three kinds of pairs are presented:

(Nat-true,Pred-true) 9 pairs with the original natural prosody vs. the original computed prosody (direct comparison).

(Nat-true,Nat-false) 32 pairs with the original natural prosody vs. a perturbed natural prosody.

(Pred-true,Pred-false) 32 pairs with the original computed prosody vs. a perturbed computed prosody.

For each pair, four choices were available:

(Nat-true,Pred-true) answer 1: natural version preferred; answer 2: computed version preferred; answer 3: both acceptable; answer 4: none acceptable.

(Nat-true,Nat-false) or (Pred-true,Pred-false) answer 1: "true" version preferred; answer 2: "false" version preferred; answer 3: both acceptable; answer 4: none acceptable.

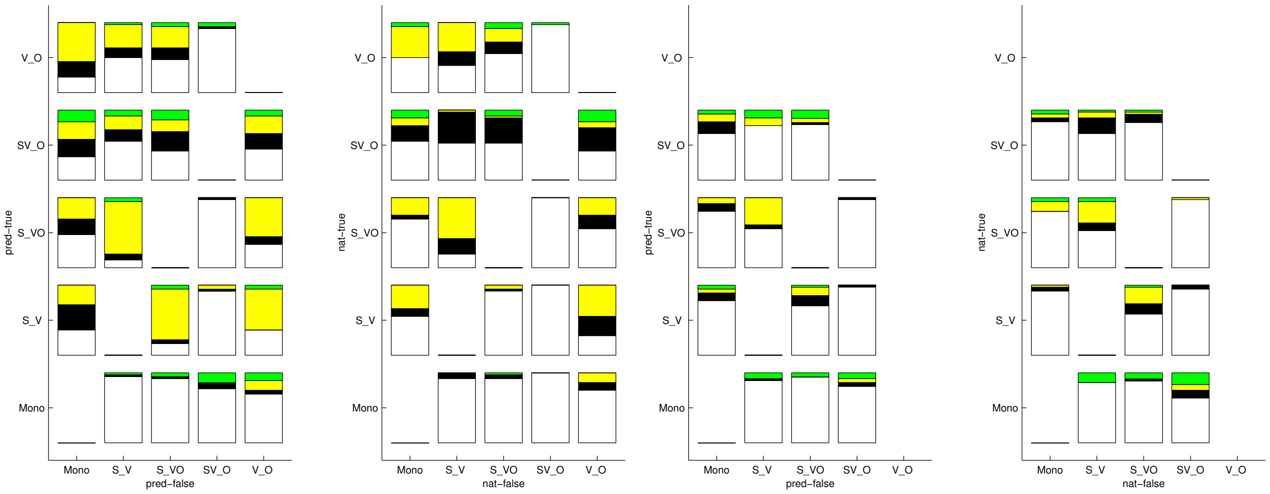
Results

The results of the experiment are shown in figures 2 and 3. The analysis of these results points out several patterns of behaviour in listeners' choices that are listed below.

Predicted versus natural prosody

- Listeners have no clear preference between natural and computed prosody (see figure 2): "natural version preferred" and "computed version preferred" have an equivalent rate of 22 % of global results. "Both acceptable" represents more than 50% of total answers.

- The most important difference is for single polysyllabic words. For 6 syllables, the natural version is preferred, whereas the reverse occurs for 8 syllables.



(a) 6 syllables

(b) 8 syllables

Figure 3: Each plot displays a confusion matrix showing preference results for (a) 6-syllable and (b) 8-syllable predicted (left) versus stylised (right) natural prosody . The ordinate is the syntactic structure of the original sentence i.e the phonetic substrate. The abscissa is the structure whose prosodic contours replace the original ones. Pairs will be labelled (true,false). White indicates the preference rate for "true" prosody, black the rate for "false" prosody, light gray indicates equivalence (both acceptable) and dark gray rejection of both stimuli (none acceptable).

The first result could be explained by the fact that the model does not take the word level into account for now: no special reduction of the duration of polysyllabic words is applied.

The second result is a consequence of the instructions given to the listeners: polysyllabic words with 8 syllables have internal morphological boundaries which were marked in the natural item: listeners thus chose the unmarked predicted version.

- Rejection rate cues the quality of the prosodic transfer to the monotonous utterances. This rate is low but we have a significant rate for the 6-syllable SV_O sentence (fig. 2-(a) and fig. 3-(a), (true,false)=(SV_O,*)).

  A careful examination of the synthetic stimuli shows an abnormal lengthening of the initial fricative [s] of the verb "scient". This was due to a strong devoicing of the following [i] which was attributed to the fricative by the automatic segmentation and which was not revised.

Common features between (Nat) and (Pred)

- The correlation between the confusion matrices (Nat) and (Pred) (figure 3) is (a) 0.844 for 6-syllable sentences and (b) 0.968 for 8-syllable sentences.
- Subjects retrieve the original prosody with greater accuracy in long sentences than in short sentences (79 % of answers for 8 syllables vs. 59 % for 6 syllables). Answer 3 "both acceptable" is more frequent for 6-syllable (22 %) than for 8-syllable sentences (8 %).
- When the original sentence is (Mono), subjects retrieve the original prosody (see pairs (Mono,*)).
- On the one hand, when a structure is competing with SV_O (see pairs (*,SV_O)), the original prosody of that structure is always chosen. On the other hand, original SV_O is highly confused with other structures for 6-syllable sentences (fig. 3-(a) (SV_O,*)).

  This shows that the boundary between V and O is rarely marked when the total number of syllables is 3 and could not fit with other prosodic phrasing.
- When a pair (S_VO,S_V) is presented, answer 3 "both acceptable" is frequent. This confirms that the V/O boundary need not to be marked.
- This remark is reinforced by the occurrence of answer 3 for (V_O,Mono).
- For 6-syllable utterances and pairs (S_V,V_O) or (V_O,S_V), answer 3 has an important occurrence rate. This shows that prosody doesn't clearly discriminate these two different boundaries.

Main differences between (Nat) and (Pred)

The global trends are the following:

- Main differences are for 6 syllables.
- Answer 3 "both acceptable" is more frequent for predicted pairs than for natural ones.
- A more precise analysis leads to the following remark: the answer 3 rate is high for computed (S_V,S_VO) and very low for natural (S_V,S_VO). This shows that our model reduces the V/O boundaries prosodic mark for a three syllable VO group. This behaviour was clearly perceptible in the corpus analysis but in the selected natural item, the boundary was realised prosodically.

Number of syllables	6						8							
Syllable location	1	2	3	4	5	6	1	2	3	4	5	6	7	8
Single word item	Mono Sentimentaliser						Mono Cytopathogénicité							
S_V item	S Des gamins			V papotaient			S Les dandys			V monopolisaient				
S_VO item	S Les gamins			V visaient		O Jean	S Les gamins			V jouent	O les mécanos			
SV_O item	S Les gars		V scient	O des rondins			S Roger		V fait	O du chocolat chaud				
V_O item	V J'adorais			O tes cheveux			not represented							

Table 1: Syntactic structures and corresponding 6 syllable length (left) and 8 syllable length (right) sentences used in the experiment. Single word, subject, verb and object are respectively labelled Mono, S, V, O. The common prosodic boundary location is placed between the third and the fourth syllable and labelled "_".

Conclusions and perspectives

In our view the way prosodic contours encode linguistic information and the way prosodic and linguistic structures are synchronised and interact are essential features of naturalness. Our methodology consists thus in evaluating the functional equivalence of predicted prosodic contours and prosodic trajectories produced in a given situation of communication.

Global results of the present experiment show a good functional equivalence between natural and synthetic prosody for a simple phrasing task. However, computed prosody sometimes fails to maintain distinctiveness. This is probably due to a lack of local prosodic indices belonging to lower levels of analysis (lexical level, morphological level…) and erased by the statistical process. This explanation has to be confirmed by an acoustical analysis of utterances proposed during the test.

This first application of our methodology to prosodic phrasing may be extended to other linguistic or phonotactic dimensions such as propositional dependency or attitudes as well as pure prosodic constraints such as speech rate or clarity.

Acknowledgements

This work has been supported by ARC-B3 initiated by AUPELF-UREF and Cost 258 "The Naturalness of Synthetic Speech".

References

Benoît, C., Grice, M. & Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences, *Speech Communication* **18**: 381–392.

Carlson, R., Granström, B. & Klatt, D. H. (1979). Some notes on the perception of temporal patterns in speech, *Proceedings of the IXth International Conference on Phonetic Sciences*, Copenhagen, Denmark, pp. 260–267.

Gibbon, D., Moore, R. & Winski, R. (1997). *Handbook of standards and resources for spoken language systems*, Mouton de Gruyter, Berlin-New York.

Lhote, E., Filleau, M. & Grange, F. (1975). Reconnaissance de patrons intonatifs, *Actes des VIèmes Journées d'Etude sur la Parole*, Toulouse, France, pp. 29–37.

Liberman, M. & Streeter, L. (1978). Use of nonsense-syllable mimicry in the study of prosodic phenomena, *Journal of the Acoustical Society of America* **63**(1): 231–233.

Lindblom, B. & Lindgren, R. (1985). Speaker-listener interaction and phonetic variation, *PERILUS IV - Publication of the Department of Linguistics* .

Morlec, Y. (1997). *Génération multiparamétrique de la prosodie du français par apprentissage automatique*, PhD thesis, Institut National Polytechnique de Grenoble, Grenoble – France. http://www.icp.inpg.fr/ ˜ morlec/_postscript/morlec_PhD_thesis.ps.gz.

Pisoni, D. B. (1997). Perception of synthetic speech, *in* J. P. H. Van Santen, R. W. Sproat, J. P. Olive & J. Hirschberg (eds), *Progress in Speech Synthesis*, Springer Verlag, New York, pp. 541–560.

Strom, V. & Widera, C. (1996). What's in the "pure" prosody, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 3, Philadelphia - USA, pp. 1497–1500.

't Hart, J., Collier, R. & Cohen, A. (1990). *A perceptual study of intonation*, Cambridge University Press, Cambridge, UK.