Keller, E., & Zellner, B. (1996). Output requirements for a high-quality speech synthesis system: The case of disambiguation. In Ch. Boitet (ed.), *Proceedings of MIDDIM-96 Seminar on Interactive Disambiguation*, pp. 300-308. Col-de-Porte, France.

Output Requirements for a High-Quality Speech Synthesis System: The Case of Disambiguation

Eric Keller
(eric.keller@imm.unil.ch)

and

Brigitte Zellner
(brigitte.zellner@imm.unil.ch)

Laboratoire d'analyse informatique de la parole (LAIP)
Faculté des Lettres, Université de Lausanne, 1015 LAUSANNE, Switzerland

Abstract

Sometimes, speakers attempt to disambiguate multiple interpretations by means of modifications of prosodic parameters, while at other times, they use various types of circumlocutions for this purpose. In view of a potential implementation of prosodic disambiguation in high-quality speech synthesis systems, the following questions arise: (1) Which types of ambiguity are open to prosodic disambiguation? (2) How are prosodic parameters modified as a function of the disambiguation attempt? (3) Can a high-quality speech synthesis system mimick prosodic disambiguation effects? We report on small pilot project performed to explore these issues.

Introduction

LAIPTTS is a high-quality text-to-speech system for French, developed at the University of Lausanne. Its creation was guided by two main objectives:

- "High Quality": Synthetic speech should resemble natural speech as much as possible, so as to permit the greatest possible ease in the comprehension of the spoken message (see Sanderman, 1996). In our system, this has been accomplished by paying exceptionally close attention to the phonetic details of speech generation, especially those related to prosody.
- "Compactness": The synthesis mechanism should require as few computational resources as possible. This has been accomplished by favoring algorithmic solutions over database solutions, and by implementing only a proximal grammar, i.e., rendering grammatical relations sensitive primarily to immediately surrounding lexical elements, and maximally, to those lying in the range of the prosodic group.

The current status of the system (autumn 1996) is as follows: the lexico-grammatical, phonological and prosodic modules are by and large completed, and the diphone output module is scheduled to be completed in the summer of 1997. Currently, we use the Mons (Belgium) MBROLA diphone output system to produce audible output[1]. The system consists of a 350k application (without interface) and uses a 2 Mb dictionary as well as a 4.8 Mb diphone database.

Like most text-to-speech systems, LAIPTTS is structured into four main modules (Figure 1). The first module takes written text and generates an annotated phonetic chain of each sentence on the basis of a dictionary and graphemo-phonetic rules. The chain is parsed into prosodic groups, and various

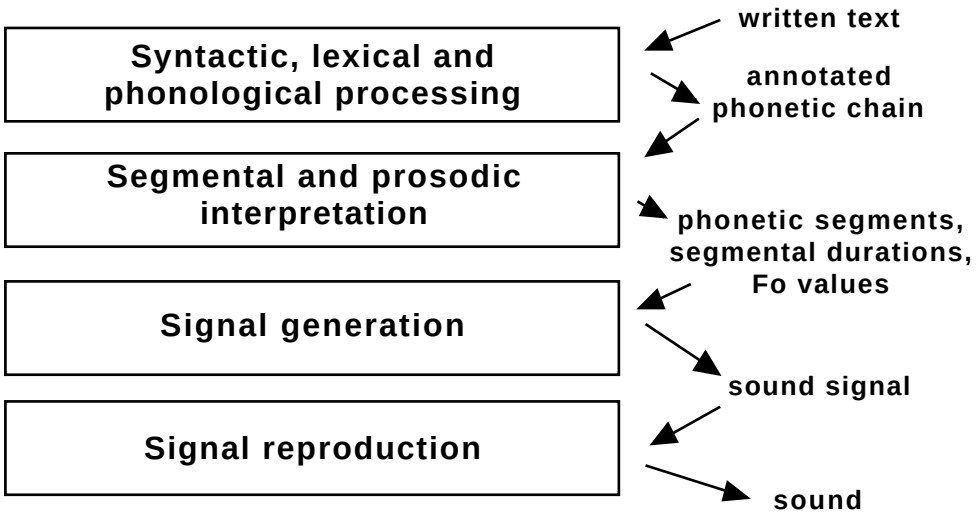
[1] The MBROLA diphone output system is produced by Thierry Dutoit <dutoit@tcts.fpms.ac.be> of Mons, Belgium. For details on MBROLA, please see http://tcts.fpms.ac.be/synthesis.

phonological rules (liaison, chaining, schwa-handling, syllabification) are applied. In the next module, durations and Fo values are generated. These values are combined with diphone segments in the subsequent module, after which the entire signal is reproduced. Calculations for the second and following sentences in a text are performed in parallel with sound outputting of the preceding sentence. This provides real-time synthesis performance on unrestricted text on most high-entry personal computers (PowerPC, Pentium-level).

Key Output Requirements

Implementation details of a synthesis system can be summarized in terms of its output requirements. To place the implementation of prosodic disambiguation strategies into its appropriate context, we shall briefly review the three key output requirements related to prosody, as well as the implementations we have chosen to meet these requirements.

Prosodic grouping: The system must perform word grouping in the way human speakers do. The placement of group marks, as well as the implementation of pauses and group-final lengthening, depends on this process. In order to meet this requirement, we have developed a psycholinguistic algorithm oriented towards timing which is largely inspired by, and builds upon, the large body of research on psycholinguistic indicators of prosodic grouping by Grosjean and colleagues (Gee & Grosjean, 1983; Monnin & Grosjean, 1993; Keller *et al*. 1993; Zellner, 1996) (Figure 2). For a first cut, the distinction of two levels (major groups, minor groups) appears sufficient for the purpose of predicting timing. For reasons detailed in Zellner (1996), psycholinguistic grouping algorithms appear to provide better predictions for overall prosodic grouping in French than do syntactically-based algorithms.



Written text: «Bonjour, comment allez-vous?»
Phonetic chain: b^ / ZuR+, 'kO / m@+ 'ta / le+ vu-?
Segments/segmental durations/Fo:

Segment	Duration	Fo
b /b/	71	107
on /^/	117	107
j /Z/	68	109
ou /u/	131	126
r /R/	54	145

Figure 1. Overview of LAIPTTS. Top: general schema. Bottom: sample sentence, with written text input, phonetic chain intermediate output, and input to the signal generation module.

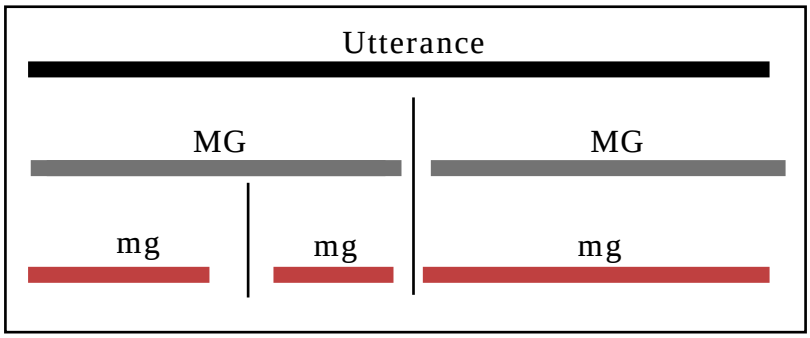


Figure 2. Hierarchic temporal structure of an utterance. An utterance is composed of temporal Major Groups (MG) which are subdivided into temporal minor groups (mg). From Zellner (1996).

Timing: The system must implement segment durations in accordance with the segmental, syllabic and phrase-level environment. To meet this requirement, we have developed a statistical model incorporating segmental, syllabic and phrase-level predictors. This work was based on an extensive review of factors influencing timing in French, and uses methodology which is largely similar to that used by Campbell (Campbell, 1992; Keller & Zellner, 1993, 1996). Correlations between measured and predicted durations for a fast-paced, but natural-sounding reading 100 sentences of French text gave, at the segmental level r = .72, and at the syllabic level r = .85, $N_{segments}$ = 4533.

The main factors of this predictive, statistical model were the following. Segmental predictors: the durational class of the predicted and the surroundingsounds. Syllabic predictors: the number of segments in utterance, the word and the syllable; the presence of a grammatical or lexical word; and the presence of an accented syllable. Phrase-level predictors: the proximity of a prosodic phrase marker or the end of the utterance.

Fundamental Frequency (Fo, intonation): The system must capture natural-sounding intonation patterns. To meet this requirement, a doctoral student of our group (Stefan Werner) adapted Fujisaki-modelling of Fo to French (Fujisaki, 1988, 1992; Werner, 1995, 1996). Specifically, he developed a neural network-based technique for deriving parameters adequate for French continuous text. It is recalled that Fujisaki modelling specifies two main parameters. (1) A phrase-level parameter, which is a rise-decline for the entire utterance. Our implementation of this first factor follows traditional Fujisaki modelling. (2) A set of syllabic parameters, which are syllable-length rise-declines superimposed on the phrase-level rise-decline. In traditional Fujisaki modelling, these rise-declines are only given for accented syllables. In French, we have found it more propicious to implement a rise-decline for each syllable, with some rises (those for accented syllables) slightly higher than others. In addition, we have defined a third set of parameters, (3) a modulation of the phrase-final contour for the specific type of sentence. This has been implemented as specific Fo modulations for the final two syllables, choses according to the sentence's declarative, paragraph-final, interrogative and exclamatory status.

The Case of Ambiguity: A Pilot Study

When considering prosodic disambiguation strategies for inclusion in a TTS system, an initial question concerns the perceptual distinctiveness. As many of the papers in this volume document, ambiguity is pervasive in common dialog situations, and speakers use a variety of circumlocutory and prosodic strategies to disambiguate sentences. Since only prosodic strategies are of interest here, we must try to estimate which sentences are open to prosodic disambiguation and which are not. Furthermore, speakers may *think* that they have disambiguated a given sentence by applying distinctive prosody, but in fact, not all sentences produced in this manner are perceived distinctly by listeners. To obtain an initial measure of factors affecting the choice and the implementation

of a prosodic disambiguation strategy, we have run a small pilot study.

Methodology. The following experimental procedure was applied: We obtained eight ambiguous sentences from two colleagues[2]. These were read by a native speaker of French. The sentences were pronounced slowly and *in a manner designed, as much as possible, to disambiguate the sentences*. Each sentence and each sense was recorded twice (i.e., for sentences with two interpretations, this gave four recordings). Recordings were made at 16 bits, 22 kHz directly into a computer and were rearranged in random fashion. Five subjects, all students or research assistants at the University of Lausanne and native speakers of French, volunteered for the 15-minute perception test. For each recorded sentence, they were presented written interpretations of the sentences. E.g., for the ambiguous sentence «Il atteint la grange et la ferme» ("He reaches the barn and the farm/and closes it"), they were given the question «Atteint-il la ferme ou ferme-t-il la grange?» ("Does he reach the farm or does he close the barn?"). Subjects were asked to choose the correct interpretation, and they were also asked to indicate their level of confidence that they had chosen the correct interpretation. The sentences and the perception results are given in Table I and in Figure 3, and the results on the confidence levels are shown in Figure 4.

Results. It can be seen that certain disambiguations are easily communicated by prosodic means (e.g., 1a, 1b), while many others are not communicated reliably. All listeners chose the correct interpretation of utterance 1 for each of the four presentations, but no other sentence provided such good results. Sentence 1 can thus be considered to have been disambiguated on the basis of prosodic effects.

It can also be seen that certain interpretations are heavily favored over others. For example, interpretation (a) was heavily favored over interpretation (b) in the case of sentence 5, and interpretation (b) was heavily favored over interpretation (a) in the case of sentence 7. An examination of the patterns of favored interpretation suggests that in performing this rather difficult perception experiment, subjects were probably guided by expectations grounded in previous experiences with similar grammatical contexts. For example, sentence 5 may well be a case of grammatical expectation, since the use of the gerund in «*Devant* cette somme, il hésite.» ("He hesitated *as he owed* this sum of money. ") is probably rare in current-day use of French. Similarly, the favored interpretation of «de» as "about" in «Il parle *de* l'école de cuisine lyonnaise.» ("He speaks *about* the Lyonnais school [of haute cuisine]. ") may be quite directly related to the high frequency of occurrence of «parler de» in the sense of «speak about» in current-day spoken French.

The examination of the confidence levels showed that high confidence was associated with a somewhat higher percentage of correct answers, and that low confidence was associated with a lower chance of correct answers. Subjects apparently have a certain sense of whether or not they have correctly interpreted the speaker's attempts at disambiguation. The particularly high percentage of incorrect choices associated with low confidence is statistically due to the choice of favored, but incorrect, interpretations mentioned in the previous paragraph.

Conclusions from the pilot study. It can be seen that some disambiguations are more easily performed by prosodic means than others. This has an important consequence for the design of automatic disambiguation systems, in that sentences open to prosodic disambiguation must be distinguished from those that require other means of disambiguation, such as circumlocutions, before prosodic disambiguation is even attempted.

The distinction between "prosodically disambiguatable" and "prosodically non-disambiguatable" sentences probably implicates a number of parameters, of which frequency of use appears to be an important factor to consider. Other factors may also be involved, such as the type of ambiguity, and the presence of pragmatic and semantic disambiguation contexts. If one interpretation is much less likely than another in current-day spoken use, a prosodic disambiguation strategy may well be insufficient.

---

[2] Thanks are expressed to Geneviève Caelen and Hervé Blanchon (Laboratoire CLIPS, University of Grenoble) for providing these sentences.

Table I: Test Sentences and Peception Results

Heterosemantic - Heterographic - Homophonic

(Difference in meaning, difference in spelling, similarity in pronunciation)

	French sentence	English interpretation	% correct[3]
1	a. Il a donné ce *chandail* à son cousin.	He gave this sweater to his cousin.	100
	b. Il a donné ce *champ d'ail* à son cousin.	He gave this garlic field to his cousin.	100
2	a. La poutre *faîtière* a sans doute été placée.	The *main* support beam has no doubt been put in place.	70
	b. La poutre *faite hier* a sans doute été placée.	The support beam *made yesterday* has no doubt been put in place.	80

Heterosemantic - Homographic - Homophonic

(Difference in meaning, identity in spelling, similarity in pronunciation)

	French sentence	English interpretation	% correct
3	a. Il atteint la grange et *la ferme*.	He reaches the barn and *the farm*.	60
	b. Il atteint la grange et *la ferme*.	He reaches the barn and *closes it*.	50
4	a. Un *savant compromis* a été arrêté hier.	A *wise compromise* was reached yesterday.	50
	b. Un *savant compromis* a été arrêté hier.	A *compromised scholar* was arrested yesterday.	90
5	a. *Devant* cette somme, il hésite.	*Confronted with* this sum of money, he hesitated.	100
	b. *Devant* cette somme, il hésite.	He hesitated *as he owed* this sum of money.	10
6	a. Il prend des cahiers et des classeurs *noirs*.	He takes *black* notebooks and *black* folders.	20
	b. Il prend des cahiers et des classeurs *noirs*.	He takes notebooks and *black* folders.	60
7	a. Il parle *de* l'école de cuisine lyonnaise.	He speaks *from* the Lyonnais school of haute cuisine.	0
	b. Il parle *de* l'école de cuisine lyonnaise.	He speaks *about* the Lyonnais school of haute cuisine.	90
8	a. *Marie* voit l'homme dans le parc *avec un téléscope*.	Mary, *using a telescope*, sees the man in the park.	40
	b. *Marie* voit l'homme *dans le parc* avec un téléscope.	Mary, *who is in the park*, sees the man who uses a telescope.	10
	c. *Marie* voit l'homme *dans le parc avec un téléscope*.	Mary, *who is in the park and uses the telescope*, sees the man.	10
	d. Marie voit *l'homme dans le parc avec un téléscope*.	Mary sees the man *who is in the park and uses the telescope*.	50
	e. **Marie** voit *l'homme* **dans le parc** *avec un téléscope*.	**In the park, Mary** sees the *man who uses the telescope*.	10

[3] Percent correct perceptions in pilot test.

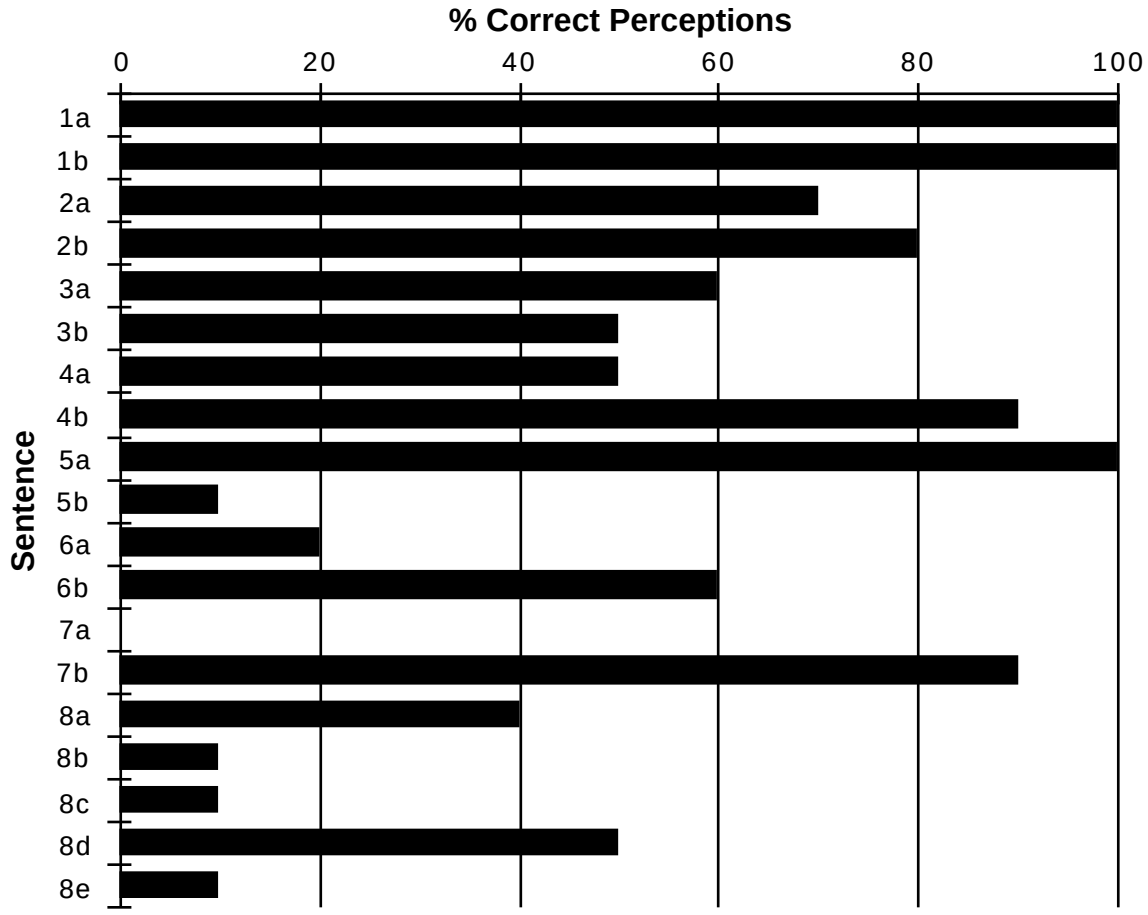


Figure 3. Percent correct perceptions on for the eight sentences of the pilot test (5 subjects and 2 presentations of each interpretation of each sentence per subject).

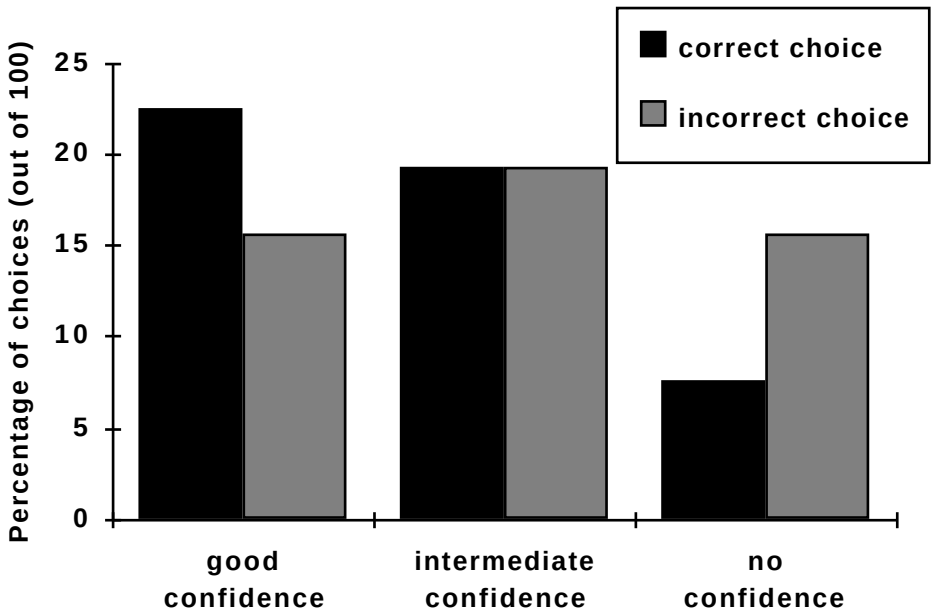


Figure 4. Correct and incorrect choices, subdivided according to confidence level.

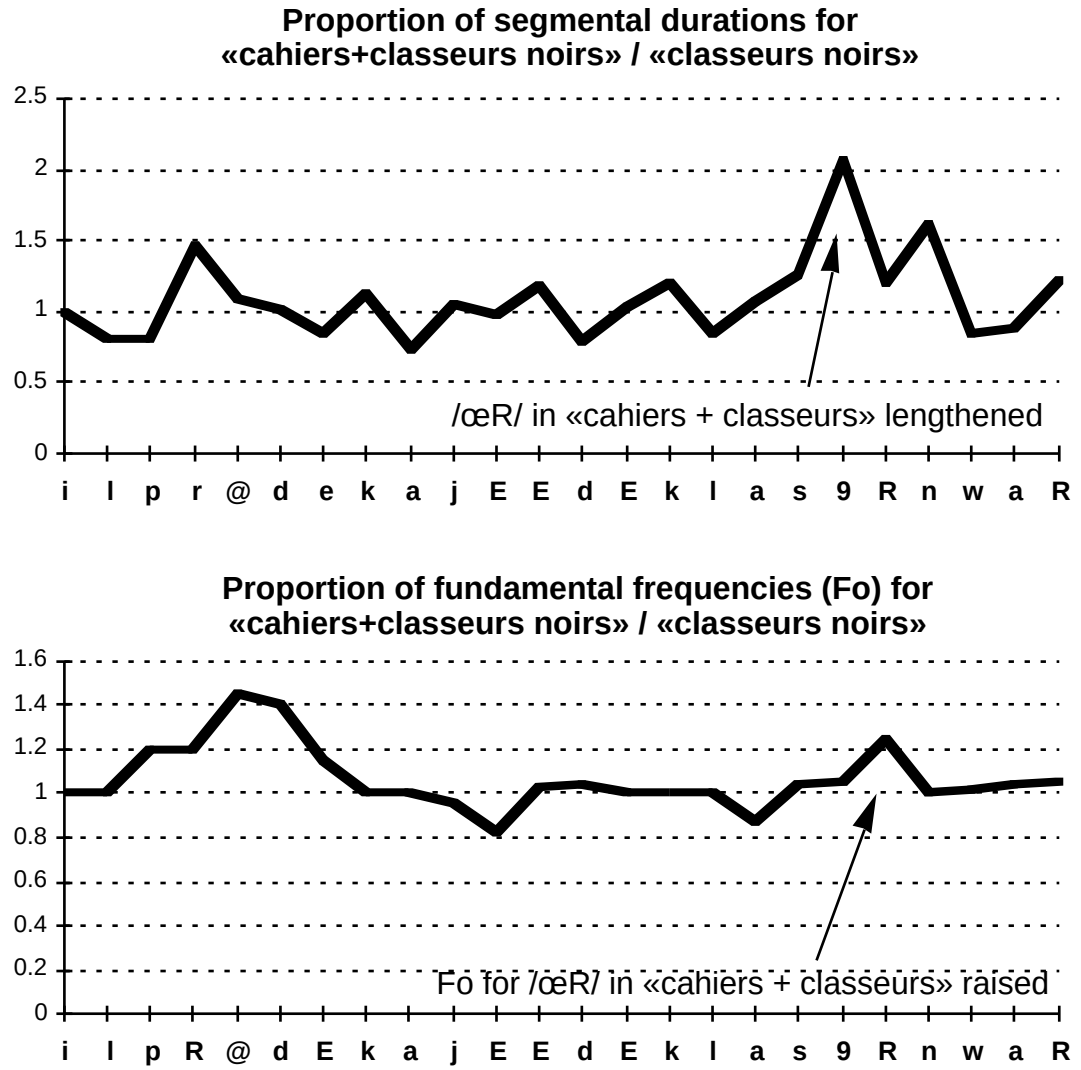


Figure 5. Proportions of segmental durations and Fo for the two interpretations of the utterance «Il prend des cahiers et des classeurs noirs.». Transcriptions on the x scale are in a 7-bit phonetic transcription. It can be seen that the syllable /πR/ of «cahiers et class*eu*rs noirs» (the /9R/) is longer and has a somewhat higher Fo than the /πR/ of «class*eu*rs noirs». In order to obtain a continuous line for Fo, unvoiced segments were attributed the Fo value of the subsequent voiced segment.

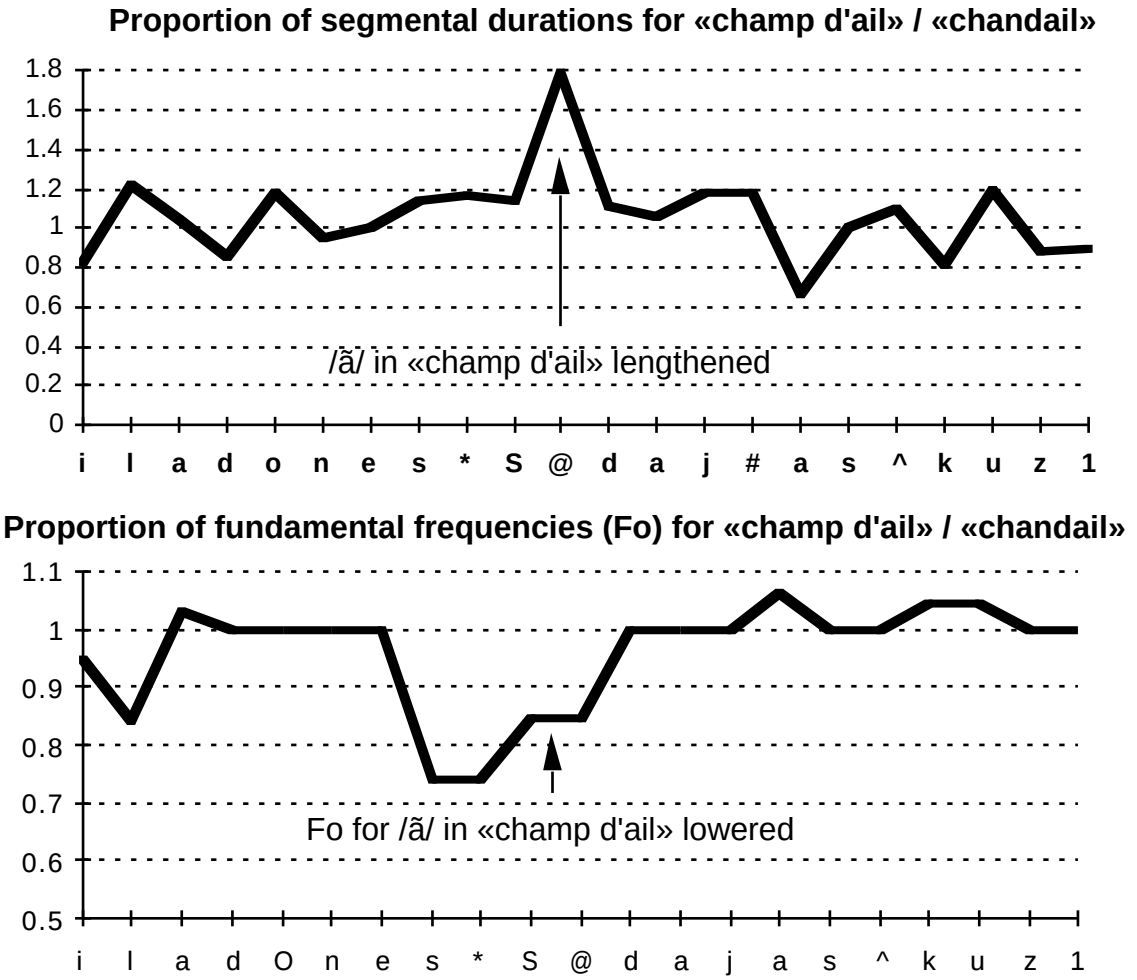


Figure 6. Proportions of segmental durations and Fo for the two interpretations of the utterance «Il a donné son champ d'ail/chandail à son cousin.». Transcriptions on the x-scale are in an 7-bit phonetic transcription. It can be seen that the /a$/ of «champ» (the /@/) is longer than the /a$/ of «chandail», at the same time as the /a$/ of «champ» (the /@/) is is part of a lower Fo-sequence than the /a$/ of «chandail» .

Prosodic Parameters for Disambiguation

In this section, we turn torwards a more detailed examination of prosodic modifications used in disambiguation. Since our pilot study furnished only one clear example of a successful prosodic disambiguation, we compare the sentence in question ([1] «Il a donné son champ d'ail/chandail à son cousin.») with another sentence for which there was no evidence that the attempt to disambiguate the sentence by prosodic means had succeeded ([6] «Il prend des cahiers et des classeurs noirs.».)

For each of the two readings of the utterance we measured, and compared by means of proportions, durations and fundamental frequency (Fo) for each segment (sound) of the utterance. These two prosodic parameters are commonly examined and manipulated in TTS systems. Proportions illustrate well the differences found in each sentence pair, since they neglect non-distinctive Fo and timing details and graphically emphasize distinctions between paired readings. The results are shown in figures 5 and 6. Some interesting divergences can be observed.

Despite the fact that subjects were not able to reliably distinguish the two interpretations of the sentence «Il prend des cahiers et des classeurs noirs.» ("He takes *black* notebooks and *black* folders" vs. "He takes notebooks and *black* folders."), some clear prosodic differences were apparent in the readings (Figure 5). For the interpretation in which the qualifier "noir" was associated with two nouns, the speaker slowed down

noticeably on the last syllable of «classeur», just before enunciating the qualifier «noir». A somewhat raised Fo is also in evidence at the very end of that syllable. We notice therefore a pairing of a lengthened duration with a raised Fo.

The opposite pattern of *divergence* between duration and Fo prevails in the case of the successful prosodic disambiguation (Figure 6). The vowel of the word «champ» (part of the two-word group «champ d'ail») is nearly twice as long as the comparable vowel of the first syllable of the word «chandail». Also, the word «champ» is part of a two-word group («ce champ») for which the Fo undergoes considerable lowering. It is likely that the successful prosodic disambiguation of this sentence was based on the combination of these two factors.

This is also suggested by the subsequent implementation of the prosodic disambiguation stragegy. We rendered our prosodic interpretation module sensitive to a special mark for "explicitness", by translating the mark into considerable lengthening associated with a lowered Fo. Casual listening to sentences generated with and without the explicitness mark easily confirmed the distinctive value of these prosodic modifications. Just as Sanderman demonstrated by means of an experimental study (Sanderman, 1996), our experiment suggests that certain combinations of prosodic parameters can favor perceptual disambiguation.

Conclusions

Our study suggests that it is important to be able to predict which ambiguities can be successfully disambiguated by prosodic means and which cannot. There is no use in attempting prosodic disambiguation if listeners cannot, or are not sufficiently prepared to, use prosodic disambiguation cues. It also suggests that if two different interpretations are indeed possible according to common spoken-language use, a prosodic disambiguation strategy may well go a considerable distance towards successful disambiguation. Finally we found that if prosodic modifications are clearly identified, current high-quality speech synthesis systems have no difficulty in implementing and transmitting the disambiguation to listeners.

Unfortunately, the small scope of the present study limits the conclusions concerning two key issues, i.e., (a) what determines whether a prosodic strategy is likely to be successful or not, and (b) exactly what the various prosodic modifications may be. Fo / duration modifications of the type described here are unlikely to be the only prosodic implementation strategy for disambiguation. Other prominent candidates would appear to be accelerations over certain stretches of speech to suggest particular semantic coherence, and punctuated Fo rises to indicate the semantic foci of a given sentence. A full inventory of such indices is clearly called for.

References

Campbell, W.N. (1992). Syllable-based segmental duration. *Talking Machines. Theories, Models, and Designs* (pp. 211-224). Elsevier Science Publishers.

Fujisaki, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In Fujimura, O. (Ed.), *Vocal physiology: Voice production, mechanisms and functions.* New York: Raven.

Gee, J. P., Grosjean, F. (1983). Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology 15*. 411-458.

Keller, E., Zellner, B., Werner, S., & Blanchoud, N. (1993). The Prediction of Prosodic Timing: Rules for Final Syllable Lengthening in French. *Proceedings,* ESCA Workshop on Prosody (pp. 212-215). Lund, Sweden.

Keller, E., & Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics*, 17, University of York. 53-75.

Monnin, P., & Grosjean, F. (1993). Les structures de performance en français: caractérisation et prédiction. *L'Année Psychologique, 93.* 9-30.

Sanderman, A. (1996). *Prosodic Phrasing. Production, Perception, Acceptability and Comprehension*. Thesis, Technische Universiteit, Eindhoven, The Netherlands.

Werner, S. (1995). Use of a neural network for parameter optimization in Fujisaki models of intonation. Poster presented at *NODALIDA 95 (Nordic Conference for Computational Linguistics)*, Helsinki, July 29, 1995.

Werner, S. (1996). The automatic generation of French declarative intonation using Fujisaki modelling. Paper presented at the *Séminaire de DEA: Synthèse de la parole, Universités Paris 7 & Paris 3.* Paris, February, 1996.

Zellner, B. (1996). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée: La communication parlée. 1.* 7-23. Paris.