Traitement des incises en français : capture automatique et modèle prosodique

Philippe Boula de Mareüil & Estelle Maillebuau

Elan TTS & LIMSI-CNRS

4 rue Jean Rodier – F-31400 Toulouse
Tél. : ++33 (0)561 36 07 77 – Fax : ++33 (0)561 36 07 70
Mél : pboula@elan.fr, mestelle@wanadoo.fr
http://www.elan.fr

BP 133 – F-91403 Orsay CEDEX
++ 33 (0)169 85 81 19 – Fax : +33 (0)5169 85 80 88
Philippe.Boula.de.Mareuil@limsi.fr
http://www.limsi.fr

ABSTRACT

Parentheticals in French are investigated, in order to assign them a specific prosody in text-to-speech synthesis. On the basis of lexical-syntactic and punctuational criteria, we show that it is possible to detect them automatically by a regular grammar, with an *f*-measure of more than 92%, on a corpus of more than 3,000 newspaper sentences, containing about 200 parentheticals. A 20% reduction of pitch range and of average pitch, as well as a 2 dB reduction of energy (respectively, a 10% reduction of pitch range) may then be applied to non-final (respectively final) parentheticals, which reflects observations made on a professional female speaker.

1. INTRODUCTION

Cet article traite de la capture automatique des incises en français et de l'application à celles-ci d'un modèle prosodique spécifique pour la synthèse de la parole à partir du texte (par diphones). L'incise est une sorte de décrochement, souligné à l'oral par des pauses, qui rompt la construction de la phrase : on peut adjoindre des indications annexes telles que le nom de la personne à laquelle on parle ou un commentaire comme dans *il viendra, j'imagine, demain*. Ces vocatifs, informations secondaires, appositions ou explications ne modifient pas la mélodie des autres éléments de la phrase. En fin d'énoncé, on reste sur le même niveau intonatif : ainsi Delattre fait-il état d'une parenthèse basse (que l'on rencontre dans *tu l'as manqué, ton train*) et d'une parenthèse haute (ou « écho », correspondant à *tu l'as manqué, ton train* ?) [Del66]. Pour désigner ces unités, Morel parle de « rupture vers le bas » [Mor98] et Mertens d' « appendice », équivalent à un « groupe intonatif » [Mer87] : en position non finale, l'incise se place sur un registre plateau (peu modulé), plus bas que le membre de phrase qui la précède ; elle se prononce ordinairement avec une diminution de l'intensité de la voix. L'intonation reprend ensuite au niveau où elle a été laissée, ce qui permet de souligner l'autonomie de la proposition enchâssée [Fag02]. L'étude menée dans des conditions étroitement contrôlées sur quatre locuteurs francophones montre un changement de hauteur par rapport à l'ensemble de la phrase, même si les corrélats intonatifs de l'incise (le degré d'abaissement des pics et des vallées de fréquence fondamentale) dépendent du sujet parlant. Les incises représentent donc un phénomène intéressant pour la théorie linguistique, pragmatique (où elles sont associées aux actes de langage illocutoires), syntagmatique (où elles permettent des arbres avec branches croisées) et phonologique [Fag01].

Il n'existe pas de consensus sur la définition des incises — ou « phrases extraposées » [Wun87]. Nous adoptons pour notre part celle restrictive mais précise de Grevisse, qui distingue *incidente* (« sous-phrase insérée à l'intérieur d'une autre phrase sans y jouer le rôle de sujet ou de complément ») et *incises* : « incidentes particulières indiquant qu'on rapporte les paroles ou les pensées de quelqu'un. Elles sont placées à l'intérieur de la citation ou à la fin de celle-ci. Le sujet est placé après le verbe » [Gre80].

La section suivante est consacrée à l'élaboration d'un corpus d'incises et d'une liste de formes verbales susceptibles de se trouver en attaque d'incise. La détection automatique des incises par une simple grammaire régulière est ensuite décrite (section 3). Section 4, la prosodie d'un ensemble de phrases avec incises est enfin analysée, et des règles appliquées à la synthèse de la parole à partir du texte sont proposées.

2. CORPUS

2.1 Incises

D'un corpus de plus de 3 000 phrases phonétiquement équilibrées du journal *Le Monde* [Lam91], nous avons extrait 196 phrases (de près de 18 mots en moyenne) contenant une incise avec un verbe de discours. Nous avons exclu les incises introduites par *comme dit…* et *comme le dit…*, où le pronom sujet ne peut être inversé : *comme (le) dit un tel* est correct, mais **comme (le) dit-il* ne l'est pas. Nous avons également exclu les phrases où l'incise en *dirait-on* est totalement débranchée du sens et du fonctionnement du verbe *dire* : il ne s'agit plus de paroles rapportées en tant que telles [Bla89].

Nous avons 25 % d'incises en milieu de phrase et 75 % d'incises en fin de phrase. Les premières ne comprennent, à une exception près, aucun circonstant, apposition ni proposition subordonnée ; mais en fin de phrase, on a près de 18 % d'incises complexes. Exemple :

> " Aujourd'hui, l'argent et la force dominent ", commente l'avocat Yukio Yamanuchi, qui pendant plusieurs années fut le conseiller juridique du Yamaguchi-gumi.

En milieu de phrase, on observe des patrons très variés : aussi bien *"texte", incise, "texte"* que *"texte, incise texte"* pour l'emploi des guillemets, aussi bien *sujet incise verbe objet* que *sujet verbe incise objet* pour le schéma syntaxique. Les sujets des incises sont tous postposés. Quant aux verbes des incises, ils sont à 95 % à la 3e personne du singulier : présent (71 %, *dit* étant compté comme tel), passé composé (17 %), imparfait (5 %), passé simple (3 %), plus-que-parfait (2 %), futur (1 %). Bien sûr, ces chiffres sont très dépendants du corpus.

2.2 Verbes de discours

Nous avons répertorié près de 300 verbes de discours comme *dire.* Nous EN avons déduit les formes conjuguées à partir du dictionnaire électronique MULTEXT [Ide94], qui compte plus de 300 000 formes fléchies (35 000 lemmes). Nous n'avons dans le programme décrit ci-dessous, pour optimiser le temps de calcul, retenu que les 3es personnes et les temps du présent et du passé de l'indicatif : on l'a vu, ce sont les plus fréquents dans les

incises (de discours rapportés). En revanche, nous avons complété la liste pour les formes qui ne figuraient pas dans MULTEXT (ex. *réprimander*). Le tableau 1 fournit d'autres exemples.

Tableau 1 : exemples de verbes de discours (ceux en gras sont représentés dans *Le Monde*).

admettre, comploter, cracher, croire, décréter, démentir, **écrire**, espérer, exiger, gémir, grommeler, montrer, **noter**, prouver, **reconnaître**, relativiser, **relever**, ricaner, s'apaiser, se plaindre, se soucier, **souligner**, soumettre, **suggérer**, supplier, **téléphoner**

On peut rencontrer des verbes habituellement intransitifs (ex. *crier*), qui indiquent de quelle manière le discours a été fait ou quel jugement porte le narrateur sur les paroles rapportées. La liste n'est pas pour autant exhaustive : les auteurs sont libres de ne pas se cantonner à ces verbes. Ex. cité par Grevisse, emprunté à *Princesse folle* de Billy :

> Du secours ! sursauta la visiteuse.

3. CAPTURE AUTOMATIQUE

3.1 Expression régulière et heuristique

Les incises sont repérées automatiquement à l'aide de critères lexico-syntaxiques et ponctuationnels : nous procédons en cela comme [Lib92] pour l'anglais. Certes on ne peut pas s'appuyer sur les guillemets de façon efficace : ils sont souvent oubliés ou mal placés ; de surcroît, l'incise elle-même peut en contenir. Exemple :

> "Cette hypothèse est abandonnée ", déclarait, le 27 juin dernier, Mr François Kourilsky, directeur général du CNRS, en présentant son " programme de modernisation".

Notons également la phrase ci-dessous, où nous avons respecté (contrairement au journaliste) les conventions typographiques voulant qu'à l'intérieur de la citation entre chevrons soient enchâssés des guillemets anglais :

> « Que les agents alternent pays "facile" et pays "difficile", expose Hugues Capet, secrétaire général du syndicat CFDT. »

Mais d'autres signes typographiques peuvent déterminer des « frontières » et être considérés comme des délimiteurs d'incise : le point d'exclamation, par exemple. Nous avons mis au point une expression régulière ainsi qu'une heuristique pour capturer les syntagmes nominaux (`SN`). Outre les séparateurs tels que les espaces (`[ \t\n]`), les classes suivantes sont utilisées : verbe personnel (`Verbe`, pour les verbes de dire conjugués), auxiliaire (`Auxiliaire`, auquel *fait* et *devait* ont été rattachés, pour rendre compte de cas comme *fait-elle remarquer*), participe passé (`Ppasse`), infinitif (`Infinitif`), adverbe (`Adverbe`, y compris les locutions adverbiales et les formes en *–ment*), pronoms sujet et datif (`PronomS` et `PronomD`).

Il est prévu que si le sujet est un pronom personnel de la 3e personne du singulier, un « *t* euphonique » vienne s'intercaler avec des traits d'union entre le sujet et un verbe terminé par une voyelle, ou bien après l'auxiliaire si le verbe est au passé composé (ex. *a-t-il affirmé*). Une incise peut également commencer par un pronom (ex. *se dit-il*). Précisons enfin que, faute d'analyse plus riche, nos `SN` peuvent être formés de n'importe quel caractère sauf de signes de ponctuation, et acceptent un point précédé d'une seule lettre majuscule (ex. *M. Martin, K. Botos*).

Les adverbes se placent entre le verbe et le sujet, si celui-ci n'est pas un pronom personnel, ou après le sujet. D'autres éléments peuvent terminer l'incise, c'est pourquoi nous avons ajouté un `SN` optionnel à la fin des patrons avec pronom personnel comme sujet — les autres ayant déjà un `SN` obligatoire. Le tableau 2 rassemble les quatre patrons de l'expression régulière, selon que le verbe de discours est à un temps simple ou composé, et selon que le sujet est un pronom personnel ou un syntagme nominal. La syntaxe est celle de `flex`, outil de manipulation de chaînes de caractères permettant de définir de manière déclarative et compacte des patrons :

- \ banalise les caractères réservés comme le tiret - ;
- | codifie la disjonction « ou », les parenthèses () bornant le regroupement ;
- [] marque un ensemble de caractères (ISO 8859-1) ;
- l'étoile de Kleene * (resp. +) signifie au moins 0 (resp. 1) occurrence de ce qui précède ;
- ? représente l'optionalité ;
- {} appelle les motifs nommés dans les définitions, par exemple `PronomD`.

La méthode envisagée permet de ne jamais récolter plus que l'incise, puisque celle-ci s'arrête à la virgule. En revanche, les incises complexes sont récupérées tronquées ; là encore faute d'analyse complète, cette stratégie est un compromis pour repérer également les incises de milieu de phrase.

Pour éviter que les verbes de discours suivis par une subordonnée conjonctive ne soient capturés comme incises, nous avons exclu le 'q' en début de `SN`. Ceci n'est qu'une heuristique, guère motivée linguistiquement, dont le but est d'améliorer sensiblement les résultats (*cf.* tableau 3).

Tableau 2 : expression régulière pour la capture des incises. Dans la première colonne, V désigne verbe, TS temps simple, TC temps composé, S sujet, PP pronom personnel et SN syntagme nominal.

V=TS S=PP	`(,|!|\?)[ \t\n]*{PronomD}?[ \t\n]*{Verbe}(\-t)?\-{PronomS}[ \t\n]* {Adverbe}*[ \t\n]*{SN}* |`
V=TS S=SN	`(,|!|\?)[ \t\n]*{PronomD}?[ \t\n]*{Verbe}[ \t\n]+{Adverbe}*[ \t\n]*{SN}+[ \t\n]*{Adverbe}* |`
V=TC S=PP	`(,|!|\?)[ \t\n]*{PronomD}?[ \t\n]*{Auxiliaire}+(\-t)?\-{PronomS}[ \t\n]+ {Adverbe}*[ \t\n]*({Infinitif}|{Ppasse})[ \t\n]*{Adverbe}*[ \t\n]*{SN}* |`
V=TC S=SN	`(,|!|\?)[ \t\n]*{PronomD}?[ \t\n]*{Auxiliaire}+[ \t\n]+{Adverbe}*[ \t\n]*({Infinitif}|{Ppasse})[ \t\n]+{SN}+[ \t\n]*{Adverbe}*`

3.2 Résultats

Les résultats sont évalués à l'aide des taux de précision (*P*) et de rappel (*R*), que la *f*-mesure récapitule : définie comme 2 *PR* / (*P*+*R*), celle-ci est en quelque sorte une moyenne dynamique qui se rapproche de la moyenne si *P* et *R* sont proches, et décroît si *P* et *R* sont éloignés.

Tableau 3 : résultats de l'extraction automatique. Les chiffres entre parenthèses indiquent que les incises récupérées incomplètes sont comptées comme erreurs.

%	Rappel	Précision	*f*-mesure	Bruit	Silence
sans heuris.	96,9 (96,9)	87,2 (78,1)	91,8 (86,5)	12,8 (21,9)	3,1 (3,1)
avec heuris.	96,8 (96,8)	88,7 (79,1)	92,6 (87,2)	11,3 (20,9)	3,2 (3,2)

Notre programme produit, avec (resp. sans) heuristique, 212 (resp. 218) « incises » dont 27 (resp. 32) ne sont pas pertinentes. L'exemple suivant de fausse incise détectée, où le terme souligné est pris comme un verbe de discours, montre la nécessité d'une analyse plus riche, pour refuser les syntagmes qui ne sont pas accordés avec le verbe.

**, ont repris le travail vendredi 8 septembre*

Avec ou sans heuristiques, 29 incises ont été tronquées, et 7 incises n'ont pas été reconnues. L'une des causes est l'oubli de la ponctuation dans le corpus. Une autre cause est l'insertion d'appositions ou de circonstants, entre virgules, séparant le verbe du sujet : le syntagme nominal n'apparaissant qu'après, la proposition n'est pas considérée comme incise par le programme. Ex. :

" L'Europe risque d'être tiraillée, voire blackboulée " confie-t-il. (absence de virgule avant l'incise)

" N'attendez pas de grande cérémonie, a dit, résolu, un artisan de Tombouctou. "

Quoi qu'il en soit, les résultats sont satisfaisants : le rappel est bon (il y a peu de silence) et la précision convenable (même si le bruit est un peu plus élevé). L'heuristique ne modifie pratiquement pas le rappel, et a une influence positive sur la précision, ce qui améliore la *f*-mesure : elle peut donc être validée.

4. ANALYSE DE LA PROSODIE DES INCISES

Nous inscrivant dans le cadre de la synthèse de la parole à partir du texte, nous n'avons pas cherché à établir une sorte de réalisation « moyenne ». Notre analyse de la prosodie, qui est encore préliminaire, est donc mono-locuteur.

Lues par une locutrice professionnelle, 14 phrases isolées avec incises finales et 52 avec incises non finales, tirées aléatoirement du journal *Le Monde*, ont été enregistrées. Répétées si elles présentaient une erreur ou une hésitation, elles ont été numérisées à 16 bits et échantillonnées à 44,1 kHz. Leur étiquetage prosodique a été effectué en retouchant manuellement la recopie de prosodie obtenue automatiquement par un outil de recopie de prosodie développé à Elan [Bou01] : étant donné une entrée audio et un texte correspondant à ce qui est prononcé, le programme génère les caractéristiques prosodiques copiées à partir de l'original. Une application en est illustrée figure 1.

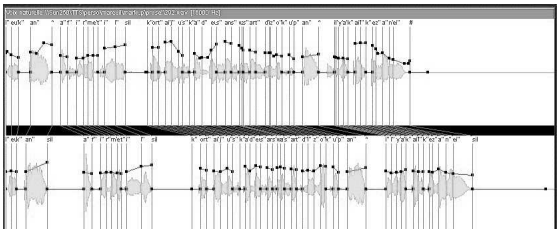

Figure 1 : exemple d'extraction de prosodie sur la phrase *Le camp, affirment-ils, comptait jusqu'à 270 occupants il y a quelques années* (en haut, voix naturelle ; en bas, voix de synthèse, où la prosodie est prédite par règles).

4.1 Incises non finales

Nous avons calculé la hauteur moyenne (des parties voisées) ainsi que le registre, défini en demi-tons comme l'écart entre les cibles maximale et minimale de fréquence fondamentale : $12 \log_2(F_0\text{max}/F_0\text{min})$. Le tableau 4 en reporte les résultats, moyennés sur les 14 phrases avec incises non finales du corpus, comparés avec ce que fournit automatiquement notre système de synthèse de la parole à partir du texte (développé au CNET, maintenant France Télécom R&D, et à Elan) pour la voix féminine du français [Bou01]. Il s'agit de la même voix, à travers la base de diphones, que celle qui a été enregistrée pour le corpus étudié.

Tableau 4 : longueur, moyenne et registre de hauteur pour les incises non finales, donnés par la voix naturelle et celle de synthèse, sur le début, le milieu (c'est-à-dire l'incise) et la fin de phrase.

hauteur		début	incise	fin
voix naturelle	moyenne (Hz)	221	180	211
	registre (demi-tons)	14	10	18
voix de synthèse	moyenne (Hz)	182	179	179
	registre (demi-tons)	10	8,5	10
nombre de phonèmes		30	14	32

Chez la locutrice, l'incise est donc prononcée avec une hauteur moyenne plus basse et un registre plus restreint que le reste de la phrase. Ceci peut s'interpréter comme une réduction de 20 % de ces deux paramètres Et si on applique cette réduction au corps de l'incise dans la voix de synthèse, on peut retrouver dans celle-ci les proportions de la voix naturelle. Par exemple, la hauteur moyenne hors incise étant de 215 Hz pour la voix naturelle et de 180 Hz pour la voix de synthèse :

$$180 \times 180 / 215 \approx (1 - 20 / 100) \times 179.$$

Si l'on regarde les syllabes finales (accentuées) suivies de pauses, on note même chez la locutrice que la montée de 8 demi-tons précédant l'incise se réduit à 2 demi-tons à la fin de l'incise dans la phrase de la figure 1 — elle passe de 6,5 à 5,5 demi-tons avec ce que prédit le système de synthèse. En revanche, nous n'observons aucune variation notable de débit ni d'allongement entre la syllabe qui précède et celle qui termine l'incise, par rapport au seuil audible [Bar87].

Nous avons également relevé une diminution de l'énergie de 40 % (2 dB) sur l'incise, par rapport au reste de la phrase chez la locutrice — alors que l'énergie n'est actuellement pas manipulée dans la synthèse. Quant aux pauses, elles sont en moyenne de 175 ms avant l'incise et de 285 ms après l'incise, chez

la locutrice. Elles sont beaucoup plus longues (près de 480 ms) dans notre système de synthèse, pour un débit de parole équivalent. Nous y reviendrons dans la section suivante.

4.2 Incises finales

De nombreuses mesures ont de même été faites sur les incises finales, mais elle n'ont dégagé chez la locutrice, entre le début de la phrase et l'incise, que des diminutions de 5 % de la hauteur moyennes et de 20 % de l'énergie (1 dB). Aucune variation de débit ni d'allongement n'est manifeste entre la syllabe qui précède et celle qui termine l'incise — par rapport aux incises non finales également. Le registre de hauteur est lui plus étroit de 10 % sur l'incise finale, par rapport aux fins de phrase de la section 4.1 (voir le tableau 5). Cette diminution peut être reportée dans la synthèse, qui sinon présente sensiblement le même registre, sur le début et sur la fin de phrase. Quant au fait que cette restriction du registre soit comparativement moindre qu'en 4.1, ce la peut s'expliquer par deux choses :

- la plus grande nécessité de distinguer les incises non finales du reste de la phrase ;
- la plus grande taille des incises finales.

Tableau 5 : longueur, moyenne et registre de hauteur pour les incises finales, donnés par la voix naturelle et celle de synthèse, sur le début de phrase et l'incise finale.

hauteur		début	fin
voix naturelle	moyenne (Hz)	215	203
	registre (demi-tons)	14	17
voix de synthèse	moyenne (Hz)	179	176
	registre (demi-tons)	9,5	9,5
nombre de phonèmes		36	26

Enfin, les pauses précédant immédiatement les incises sont en moyenne de 175 ms (comme pour les incises non finales) chez la locutrice. Elles sont presque trois fois plus longues avec la synthèse, pour un débit de parole équivalent. Dans tous les cas, il faudrait donc raccourcir la pause précédant l'incise ; mais ceci est plus délicat à exprimer à l'aide d'un langage de balise…

5. CONCLUSION

Parmi les différents langages de balise destinés à la synthèse de la parole à partir du texte que nous avons examinés (par exemple SABLE [Spr98]), aucun n'évoque le cas des incises, alors que certains prévoient de communiquer des émotions, de développer un modèle d'emphase ou d'insérer des pauses. Dans cette étude, nous avons montré qu'il est possible en français, avec de bons résultats, d'extraire automatiquement les incises afin de leur attribuer une prosodie caractéristique, de type parenthétique. D'autres applications peuvent être envisagées.

Dans le cas de notre système de synthèse de la parole à partir du texte, nous avons vu qu'une réduction de 20 % de la hauteur moyenne et du registre, et de 40 % de l'énergie, appliquée à une incise non finale, permet de refléter la réalisation d'une locutrice. Ceci peut se représenter, par exemple dans un langage comme SABLE, par :

```
<PITCH RANGE="-20%"> … </PITCH>
```

Une réduction dans des proportions deux fois moindres peut également être appliquée aux incises finales, que la détection automatique permet implicitement de distinguer.

Un travail similaire sur les titres a été ébauché, sur la base d'une grammaire régulière faisant davantage appel à des critères typographiques : il nous reste à développer un modèle prosodique adéquat. Comme pour les incises, celui-ci mériterait sans doute d'être affiné sur plus de données et par rapport au nombre de syllabes, et devrait être évalué perceptivement. Une impression de lecture plus « intelligente » est à espérer.

BIBLIOGRAPHIE

[Bar87] Bartkova K. & Sorin C. (1987), « A model of segmental duration for speech synthesis in French », *Speech Communication*, Vol. 6, pp. 245-260.

[Bla89] Blanche-Benveniste C. (1989), « Constructions verbales en incise et rection faible des verbes », *Recherches sur le français parlé,* Vol. 9, pp. 53-73.

[Bou01] Boula de Mareüil P. *et al*. (2001), « Elan Text-To-Speech : un système multilingue de synthèse de la parole à partir du texte », *Traitement Automatique des Langues,* Vol. 42, pp. 223-252.

[Del66] Delattre P. (1966), « Les dix intonations de base du français », *French Review,* Vol. 40**,** pp. 1-14.

[Fag01] Fagyal Z. (sous presse), « Prosodic boundaries in the vicinity of utterance-medial parentheticals in French », *Probus*, Special issue on the intonation of Romance languages édité par J. I. Hualde.

[Fag02] Fagyal Z. (sous presse), « Intonation in utterance-medial parentheticals and the syntax-phonology interface in French », *in* Andronis M., Ball C., Elston H. & Neuvel S. (eds.), *CLS,* Vol. 37 : The Main Session Papers from the 37th Meeting of the Chicago Linguistic Society (Vol. 1), Chicago.

[Gre80] Grevisse M. (1980), *Le Bon Usage. Grammaire du français avec des Remarques sur la langue française d'aujourd'hui*, Éditions Duculot, Paris-Gembloux.

[Ide94] Ide N. & Véronis J. (1994), «Multext: Multilingual text tools and corpora », *COLING*, Kyoto, pp. 588-592.

[Lam91] Lamel L.F., Gauvain J.-L., Eskénazi M. (1991), « BREF, a Large Vocabulary Spoken Corpus for French », *Eurospeech,* Genova, pp. 505-508.

[Lib92] Liberman M.Y. & Church K.W. (1992), « Text Analysis and Word Pronunciation in Text-to-Speech Synthesis », *in* Furui S. & Sondhi M.M. (eds.), *Advances in Signal Processing*, Marcel Dekker, New York, pp. 791-831.

[Mer87] Mertens P. (1987), *L'intonation du français*, *De la description linguistique à la reconnaissance automatique*, Doctorale dissertatie, Katholieke Universiteit Leuven, Louvain.

[Mor98] Morel M.A. & Danon-Boileau L. (1998), *Grammaire de l'intonation. L'exemple du français*, Éditions Ophrys, Paris.

[Spr98] Sproat R., Hunt A., Ostendorf M., Taylor P., Black A., Lenzo K. & Edgington M. (1998), « SABLE: a standard for TTS markup », *Third ESCA COCOSDA International Workshop on Speech Synthesis,* Jenolan Caves, pp. 27-30.

[Wun87] Wunderli P. (1987), *L'intonation des séquences extraposées en français*, Gunter Narr Verlag, Tübingen.