Ce qu'on peut faire pour la synthèse de la parole avec un peu plus de prosodie et une meilleure qualité du signal

Autor(en): **Local, John**

Objekttyp: **Article**

Zeitschrift: **Études de Lettres : revue de la Faculté des lettres de l'Université de Lausanne**

Band (Jahr): **- (1997)**

Heft 3

PDF erstellt am: **22.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-870414

CE QU'ON PEUT FAIRE POUR LA SYNTHÈSE DE LA PAROLE AVEC UN PEU PLUS DE PROSODIE ET UNE MEILLEURE QUALITÉ DU SIGNAL

Cet article propose d'aborder, de manière assez personnelle, certaines observations faites à partir de données prises dans un corpus de langue parlée pour modéliser le détail phonétique et l'information prosodique requise en synthèse de la parole. L'auteur conclut que pour améliorer la qualité actuelle de la synthèse, nous devons trouver un moyen de modéliser la «cohésion acoustique» (Hawkins & Slater, 1994) et la variabilité systématique qui caractérisent la parole naturelle et spontanée. Deux questions fondamentales sont soulevées: i. la détermination de l'étendue sur laquelle on peut modéliser les différents paramètres et ii. la formulation des interactions et interdépendances exactes entre les différents composants du ou des modèles.

1. Introduction

Je voudrais commencer par un avertissement « pour la santé du lecteur »: j'utilise délibérément le terme « prosodie » dans deux sens tout à fait distincts. Le premier sens correspond à l'usage traditionnel du terme prosodie en phonétique et en phonologie (Werner & Keller, 1994; Zellner, 1994). Dans ce sens, on considère que les caractéristiques prosodiques impliquent des paramètres non-segmentaux tels que la hauteur, le timbre et le rythme. Le deuxième sens dans lequel le terme « prosodie » sera utilisé correspond au sens donné à ce mot par les phonéticiens de « l'École de Londres », en particulier John Rupert Firth, dans leurs travaux connus sous le nom « d'analyse prosodique firthienne » ou FPA (« Firthian Prosodic Analysis », Firth, 1948). Dans ce sens, le terme « prosodie » renvoie aux catégories abstraites de la représentation phonologique. Dans les travaux de FPA, ces « prosodies » sont caractérisées par leur fonction intégrative (établissant et renforçant les liens entre les segments (les phonèmes)), et ils caractérisent la contrastivité phonologique sur

des étendues de parole plus importantes que le segment, c'est-à-dire, des étendues qui seraient considérées comme non-terminales ou supra-segmentales dans les théories classiques. Ces prosodies peuvent avoir des exposants (c'est-à-dire des interprétations phonétiques) qui iraient au delà de ce qui serait traditionnellement considéré comme segmental ; de tels exposants phonétiques sont parfois appelés « domaine-long » et servent à fournir la « colle acoustique » qui permet aux auditeurs d'analyser le flot de la parole en unités linguistiques pertinentes (voir aussi Bregman, 1990).

Malgré de récents progrès — dans les techniques de traitement du signal, dans l'analyse grammaticale, l'utilisation de modèles dans le cadre de grandes banques de données — la voix synthétique n'a pas encore atteint un niveau de qualité naturelle totalement satisfaisant, qui garantirait son utilisation et son application à grande échelle. Si les meilleures systèmes de synthèses par concaténation sont capables d'éviter certains des problèmes majeurs de la modélisation de la qualité vocale et des effets coarticulatoires locaux, il n'en demeure pas moins que la voix synthétique paraît toujours peu naturelle, et est souvent difficile à comprendre. Je voudrais démontrer que ce manque de naturel résulte d'un problème fondamental — celui du manque de cohésion intrinsèque du signal acoustique synthétisé — et que nous devons nous efforcer de résoudre ce problème sur la base de données provenant de la linguistique, si nous voulons améliorer le fonctionnement des systèmes de synthèse actuels. À mon avis, la cause fondamentale du manque de cohésion acoustique est d'abord à rechercher dans le fait que (1) nous n'avons pas réussi à modéliser les détails et la variabilité (sous-phonémiques) phonétiques (sens prosodique de Firth), et (2) nous ne savons pas bien capter les interactions entre les différentes parties des modèles que nous utilisons (par exemple, les modèles d'intonation et de durée, ainsi que les modèles de l'intonation segmentale et non-segmentale).

2. *Structure phonétique*

Il est généralement admis que la structure phonétique est une suite linéaire de segments reliés par un enchaînement non-segmental (prosodique / intonatif) globalement indépendant. Alors que la plupart des chercheurs s'accordent à dire que cette description n'est rien de plus qu'une représentation fictive mais com-

mode, c'est elle qui est à la base de presque tous les modèles phonétiques les plus influents de production, de perception et de synthèse de la parole. Néanmoins, les interdépendances entre les paramètres segmentaux, prosodiques et grammaticaux sont bien connus des phonéticiens et de tous ceux qui ont essayé de synthétiser la parole. Par exemple, la structure temporelle (le *timing*) contribue de manière cruciale à l'identité des segments et à la prosodie. Et la forme phonétique de tel ou tel type de contour intonatif en anglais, varie de manière systématique en fonction de ces « matériaux segmentaux » avec lesquels le contour est associé.

Quand ces composants (segmentaux, prosodiques et grammaticaux) sont employés pour la synthèse sous forme de modules séparés, leur commodité apparente est annulée par la nécessité de reproduire ces interdépendances, ce qui conduit souvent à des problèmes d'ordonnancement de règles, ainsi qu'à une multiplication de nouvelles règles afin de corriger les effets des règles précédentes. De plus, la parole est généralement caractérisée par des éléments relatifs — les qualités spectrales qui produisent l'impact perceptif d'une voyelle centrale dans un contexte donné peut, dans un autre environnement, résonner plutôt comme une voyelle postérieure. Ou encore, les caractéristiques de durée qui produisent des effets perceptifs spectro-temporels satisfaisants dans la deuxième syllabe non-accentuée d'un mot dissyllabique germanique (en anglais) auront presque certainement une résonance bizarre si elles sont appliquées à la deuxième syllabe non-accentuée d'un mot trisyllabique d'origine latine. (Ceci n'est pas seulement dû à la nécessité de modéliser différents taux de compression de syllabes en fonction du nombre de segments et / ou de syllabes qui augmentent dans le mot.)

Ces matières prennent une importance particulière à la lumière des résultats des recherches actuelles sur la perception et la production de la parole. Les travaux actuels montrent que les corrélats acoustiques des unités linguistiques sont « de type complexe, étendus sur des tronçons relativement longs du signal — leur effet se faisant sentir simultanément sur plus d'une unité linguistique — et les unités linguistiques ne forment pas de groupes discrets » (Hawkins, 1995, voir aussi Kelly & Local, 1986)[1].

---

1. C'est probablement l'une des raisons pour lesquelles la synthèse concaténée non-uniforme constitue une amélioration par rapport aux approches diphoniques — en d'autres termes, elle permet de se rapprocher un peu plus de certains domaines de la «colle» phonético-prosodique. Cependant le fait qu'il n'y

Par ailleurs, nous disposons de plus en plus de démonstrations que la parole est riche en détails phonétiques non-phonémiques contribuant à son naturel et à sa vigueur (p. ex. Hawkins & Slater, 1994; Simpson, 1992; Manuel *et al.*, 1995). Une des conséquences des recherches récentes en audition (Summerfield & Culling, 1994; Cooke & Brown, 1994) est que nous devrions comprendre, et ensuite modéliser, les domaines d'affectation et l'étendue temporelle de la « colle » phonético-prosodique si nous souhaitons produire une synthèse qui soit cohérente d'un point de vue sonore. En effet, les résultats obtenus dans le domaine de la ségrégation des flux auditifs montrent que lorsque des sons sont liés par certaines relations temporelles et spectrales, les auditeurs les groupent en schémas cohérents. Certaines relations spectro-temporelles sont perçues comme des schémas cohérents provenant d'une source sonore unique; celles qui ne respectent pas de tels schémas sont entendues comme provenant de différentes sources sonores, c'est-à-dire des événements acoustiques sans relation.

Certaines modifications, par exemple, dans les relations temporelles ou fréquentielles peuvent changer considérablement l'impact perceptif d'un groupe de sons. Ainsi en parole, les auditeurs utilisent la continuité du Fo pour distinguer des voyelles simultanées les unes des autres (Assman & Summerfield, 1990). Le problème ne consiste donc pas seulement à obtenir les bons segments, avec la structure temporelle adéquate, et l'intonation qui convient. Il faut aussi capter les *détails de ces inter-relations*. Il est maintenant courant de constater que même si on a réussi à développer des modules qui expliquent des parties importantes de la variance dans certains domaines de modélisation (segmental, temporel, intonatif), les résultats ne sont pas si satisfaisants quand on associe ces modules en synthèse. Souvent, il n'est pas facile d'identifier l'origine du problème. Ainsi, avec une analyse syntaxique juste, un groupe de mots peut avoir une bonne résonance, tandis qu'une autre, avec la même structure grammaticale, sonnera faux.

Dans ce qui suit, je voudrais m'exprimer en tant que linguiste-phonéticien, et attirer votre attention sur l'analyse de quelques données de parole naturelle. J'essaierai ainsi d'explorer quelles sont les implications pour modéliser les détails phonétiques, la variabilité phonétique, la « colle » phonétique et les traits proso-

---

ait pas de fenêtre optimale pour la «colle» phonétique présente un défi intéressant à relever pour l'approche concaténée (par rapport aux approches traditionnelles basées sur les formants, par exemple).

diques, et de cette façon comment améliorer la qualité sonore de la parole synthétique. Je commencerai en examinant quelques aspects apparemment insignifiants du détail phonétique.

3. Détails phonétiques et variabilité légitime

Une des difficultés que nous rencontrons quand nous tentons de modéliser la parole est, qu'*a priori*, nous ne savons pas identifier quels types de portion du détail phonétique sont nécessaires. En d'autres termes, nous ne savons pas toujours, *a priori*, quelles sont les parties du signal importantes, ni à quelles fins elles servent. Trop souvent, il a été admis par commodité que nous pouvions facilement identifier ces petits éléments permettant de modéliser le message (par exemple, au moyen d'un groupe de phonèmes) et les autres détails ont été relégués au second plan. (Il va de soi que le choix des éléments jugés « essentiels » dépendra du type de théorie linguistique — quelles qu'en soient ses limites — qui sera à la base de la réalisation.) Il est par ailleurs important de rappeler qu'un paramètre phonétique donné peut, à un seul et même moment, remplir plusieurs fonctions linguistiques significatives.

Considérons la pertinence interactionnelle d'un élément de détail phonétique dans les deux extraits suivants pris d'une banque de données de conversation naturelle et spontanée (décrites dans Local *et al.*, 1986). L'élément sur lequel je voudrais attirer votre attention est l'aspiration caractérisant toutes les occlusives sourdes en position finale dans les mots "took", "back", et "toilet". Il est traditionnellement admis que l'aspiration ou non d'une occlusive sourde en position finale devrait être un cas de « variation libre » (Gimson, 1962). Cependant, d'autres mots dans ces courts extraits se terminent par des occlusives sourdes : "got", "Vincent", "that" et "Mick", et dans aucun de ces cas, l'occlusive en finale de mot n'était produite avec une aspiration audible.

McN 1.1.3
A: have you go**t** your snaps Vincen**t** tha**t** Mi**ck** took^{H}
N: no Connie's go**t** them

McN 1.2.5
P: and I say oh oh she's away round th- the back^{H}
M: aye (.) she's e(h
(
A: (gone to the toile**t**^{H}
N: all kinds of amenities I'll tell you

Pouvons-nous conclure qu'il s'agit d'un simple cas de variation libre ? Si nous voulions uniquement modéliser un contraste *lexical*, notre réponse pourrait être « oui ». Cependant, un examen plus attentif des données nous montre que les cas d'aspiration observés représentent en fait une variation systématique régulière. En effet, nous disposons de nombreuses indications que l'aspiration en finale de mot est produite de manière systématique (et est interprétée comme telle par d'autres participants) afin d'indiquer la fin d'un tour de parole. Les données d'interaction que nous possédons sur les locuteurs de Tyneside montrent certainement que cette aspiration, associée à la centralisation de la qualité des voyelles en fin de tour de parole, est un critère caractéristique signalant le passage du tour de parole.

De plus, les schémas intonatifs qui apparaissent à ces endroits de la conversation semblent avoir pour rôle principal l'encodage d'une fonction interactionnelle et grammaticale. La règle semble être la suivante : si un mot se trouve à la fin d'un tour de parole, et si ce mot se termine par une occlusive sourde, il y a aspiration — les occlusives sourdes, ailleurs, étant glottalisées de façon coordonnée[2]. Il me semble qu'une grande partie de ce qui est communément considéré comme « variation libre » dans la langue courante est en fait une variation *systématique légitime* qui encode une variété de fonctions linguistiques et que nous ignorons, à nos risques et périls, dans nos modèles de synthèse.

Considérons maintenant les observations suivantes, apparemment tout aussi insignifiantes, et qui concernent la variabilité de la labialité et de la nasalité dans *I'm* (où les deux sont accentués et inaccentués) dans des énoncés tels que : *I'm opening*, *I'm bringing*, *I'm fishing*, *I'm going*, *I'm thinking*, *I'm not bringing*, *I'm washing*. (Ces observations sont faites à partir de données expérimentales obtenues de 10 étudiants au niveau de la licence (5 de sexe masculin et 5 de sexe féminin) de l'université de York, locuteurs d'anglais britannique du sud (Local, en préparation). Chez certains de ces locuteurs (8 sur 10), nous avons trouvé des gammes importantes et régulières de variabilité dans les exposants phonétiques suivants : (nasal bilabial), (nasal labiodental),

---

2. Si nous devions modéliser le finlandais, nous verrions que les occlusives apicales (pré-accentuées) en fin de mot ont une fonction différente: l'occlusion et l'échappement d'air aspiré traduisent un contraste lexical, alors que l'apicalité est l'exposant de la consonantalité en fin de mot — en finlandais l'apicalité n'apparaît qu'en fin de mot.

(nasal vélaire), (nasal dental), (nasal alvéolaire), (approximante labio-vélaire nasalisée). Par contre, la variabilité apparaît bien plus limitée (exposants (nasal bilabial) et (nasal labiodental) seulement), chez les mêmes locuteurs, dans les unités du type *time*, *lime*, *mime*, *rhyme* produites dans des énoncés comme : *the lime opens easily*, *the lime fills the glass*, *the lime never tastes right*, *the lime goes quickly,* etc.

Une partie de la « colle » segmentale pourrait tomber à l'intérieur du mot si nous disposions d'un modèle de synthèse « gestuel » (Browman & Goldstein, 1989) capable de fournir une bonne explication de la première catégorie de variabilité en termes de « fusion inter-gestuelle ». Cependant, du moins dans son expression simple, une telle explication d'orientation articulatoire ne serait pas en mesure de prédire la gamme restreinte de variabilité de la deuxième catégorie. Ceci est probablement dû au fait que dans le deuxième cas, la variabilité ne concerne pas véritablement les gestes, ni les bruits eux-mêmes, mais qu'elle est — du moins en partie — liée au fait que la nasalité et la labialité de *I'm* assurent une fonction grammaticale (en tant que partie d'un auxiliaire), alors que dans *lime*, elles ont une fonction lexicale (pour distinguer ce mot de *line*, par exemple). Nous sommes en présence ici de systèmes de contrastes abstraits dans lesquels *I'm* est en relation, d'une part avec *you're*, *we're*, *they're* et d'autre part avec *he's*, *she's*. Autrement dit, il s'agit d'un système contrastif à trois termes : (nasalité), (centralité), (friction) — ce qui fait que le point d'articulation et la nasalité ne sont pas liés de la même façon que dans le cas lexical (voir Kelly & Local, 1986, 190-202).

Il est important de souligner ici que cette variabilité ne semble pas être asujettie au débit. Ce phénomène n'est pas seulement le résultat d'une manière de parler rapidement. Il est profondément lié à la structure linguistique de la langue.

L'implantation ou non de tels phénomènes dans un système de synthèse dépend des modèles phonétiques et linguistiques que nous mettons en œuvre. Il est évident que si nous utilisons un modèle basé sur l'information provenant du niveau grammatical, nous pouvons exercer une distinction entre les différents « m » de *I'm* et de *lime,* etc. De plus, si nous disposons d'un modèle phonétique à paramètres, comme le modèle de synthèse non-segmentale YorkTalk (Local, 1994), nous pouvons évaluer, en termes de durée, l'étendue des différents sous-composants de cette « colle ». Le timing de l'initiation de la nasalité semble différer sytématiquement dans les deux cas, tout comme la vitesse des transitions

des formants correspondants à la phase finale de la consonne. Il existe aussi des différences systématiques, dans les caractéristiques de durée de la deuxième partie de la diphtongue [ai] se produisant entre les formes grammaticales, qui pourraient être exploitées dans le contexte d'un système de synthèse offrant ce niveau de manipulation phonétique.

4. Prosodies intonatives — schémas et relations

Deux raisons au moins expliquent pourquoi des chercheurs essayent de modéliser quelques-uns de ces traits prosodiques. Premièrement, on a constaté l'absence, dans les systèmes TTS (Text-to-Speech), du détail acoustique-phonétique (souvent redondant et donc probablement utile à la perception) que l'on retrouve généralement dans la parole naturelle. Deuxièmement, on a constaté que les indices suprasegmentaux peuvent améliorer de manière importante la perception et la compréhension d'un texte.

Un de ces indices suprasegmentaux est la proéminence prosodique. Souvent appelée *accentuation* ou *accent*, celle-ci est traditionnellement vue comme une propriété suprasegmentale qui recouvre la matière segmentale. Si nous adoptons ce point de vue, il s'en suit que l'on développe des modèles parallèles pour traiter, dans un cas, par exemple, le *timing* segmental ainsi que l'intonation, et, dans un autre cas, pour analyser et traiter les influences minimes inter-modèles, ceci dans le cas où, et au moment où, de telles influences se manifestent.

Trois questions évidentes se posent. Premièrement, est-ce que ceci est une hypothèse adéquate ? Deuxièmement, dans quelle mesure ces effets « minimes » sont-ils véritablement minimes ? Et si ces effets existent, comment de tels traits prosodiques / intentionnels interagissent-ils avec les autres niveaux de structure ? Finalement, qu'est ce qui doit être modélisé, et comment ? De nombreux chercheurs (Keller *et al.*, 1995 ; Kohler, 1997 ; Van Santen & Hirschberg, 1994 ; Local & Ogden, 1997) ont montré les avantages potentiels de l'adoption d'une approche de modèles distincts, mais il nous apparaît de plus en plus que les relations reliant minimalement, la prosodie, la grammaire, le lexique et la fonction interactionnelle soient beaucoup plus entremêlées qu'on ne le croit à l'heure actuelle.

5. Interactions entre grammaire-lexique et prosodie

Il est bien connu que les contours d'intonation présentent différents schémas de co-occurrence avec différentes structures grammaticales. Mais ici, je voudrais examiner deux formes de contraintes gouvernant l'intonation, la syntaxe lexicale et le sens qui nous appellent à la prudence quand nous mettons en application des modèles traditionnels, par exemple, en traitant les contours d'intonation comme une couche « indépendante » qui est simplement superposée sur les segments et les mots.

Dans le premier exemple, il s'agira de la forme phonétique globale d'une particule minuscule mais fréquente en conversation — l'expression "oh !" (Local, 1996). Cette particule apparaît dans de nombreux contextes, mais l'un des plus fréquents est celui de la « réception de l'information », par exemple, lorsque quelqu'un communique une nouvelle à un auditeur à laquelle celui-ci répond "oh !". Je ne mentionnerai ici que deux cas intéressants : (1) les occurrences de "oh" dans les formulations d'appréciation et (2) les occurrences libres de "oh" dans l'information fournie en réponse à une question.

L'expression "oh" apparaît souvent dans sa forme libre — non-accompagnée de tout autre trait — dans la prise du tour de parole. Mais parfois elle peut prendre la forme de "oh really", "oh wow", "oh good". Le contour d'intonation peut également varier en fonction de ces formes : il peut y avoir ou ne pas y avoir d'évolution dynamique de l'intonation sur la durée de la syllabe (alors que les occurrence libres de "oh" montrent toutes une telle évolution dynamique). Le contour intonatif de l'ensemble du tour de parole peut monter ou descendre. En même temps, ces occurrences de "oh" débutent toutes par une occlusion glottale (qui est variable dans les occurrences libres) et elles présentent les traits typiques de la diphtongaison. Toutefois, lorsque la forme est : "oh lovely", "oh good", "oh wow", le ton va *toujours dans le sens descendant*.

Cependant, si les "oh" sont produits en tant que réponses à une information donnée à la suite d'une question, leur forme phonétique est plutôt différente. Ces expressions sont régulièrement prononcées avec une intonation finale descendante qui est systématiquement produite avec un coup de glotte initial. De plus, contrairement aux autres occurrences de "oh", elles peuvent se terminer par une occlusion complète de la glotte. Leur qualité vocalique est donc tout à fait distincte des autres occurrences de "oh".

Finalement, les "oh" produits en réponse à une question sont articulés en tant que monophtongues. Celles-ci ont les traits caractéristiques des vocoïdes postérieures, généralement ouvertes ou mi-ouvertes. Leur qualité oscille entre celle des voyelles cardinales 5 et 6. Dans la zone de la voyelle cardinale 6, la vocoïde est en général légèrement non-arrondie.

La seconde série de données est extraite du début d'un corpus de requêtes indirectes et autres énoncés du même genre, pris dans des enregistrements de conversations naturelles et spontanées.

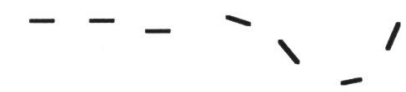
(a)

B: oo (.) it's cold in here

D: I'm sure we c'd close the doo(r
B: (please

(b)

S: hey is the window open



Sa: (sh- I-
T: (I'l- I'll get it
S: n- (.) th- can you do the door as well

(c)

A: this chair's lumpy

M: aye (.) the spring's knackered
A: oh (1.6) uh (.) what's on tonight

Les extraits (a) et (b) sont des exemples évidents de requêtes indirectes, qui sont traitées en tant que telles par les locuteurs. En ne considérant que les transcriptions de surface, on pourrait être induit à penser que l'extrait (c) ressemble sémantiquement aux extraits précédents (une plainte suivie par une réaction). Mais il apparaît clairement qu'en fait, il n'est pas traité comme une requête indirecte — ni par le locuteur A, ni par la locutrice B elle-même, car elle change de sujet au troisième tour de parole.

Ce que je vais avancer maintenant est largement basé sur des conjectures. On peut remarquer que dans (a) et (b), le contour intonatif (représenté schématiquement) a une forme clairement

montante-descendante à la fin, et que l'intonation finale y atteint un niveau plus élevé que dans les autres tour de parole. Dans l'exemple (c), malgré la présence d'un contour descendant sur "chair's", la légère élévation sur "lumpy" reste relativement basse, pris en rapport de la gamme intonative complète. Ces manipulations subtiles pourraient servir à soutenir prosodiquement les différences sémantiques caractérisant ces échanges.

Il est évident que même si cette corrélation s'avère soutenable par des observations supplémentaires, une telle interrelation entre traits prosodiques et interactionnels est plutôt subtile. Dès que nous dépassons le stade de la modélisation de la prononciation intrinsèque des phrases déclaratives, et que nous essayons de modéliser des énoncés moins réguliers de paroles telles que des questions, des réponses, des énoncés interactifs, des dialogues, etc. — il faut s'interroger sur les évolutions structurelles (s'il y en a) dont nous devrions enrichir nos modèles. Les travaux que nous avons effectués sur l'analyse de la conversation continue, font ressortir des traits phonétiques et des associations régulières entre les traits phonétiques et les niveaux linguistiques qui ne semblent pas avoir été identifiés dans les données des laboratoires. Ces données de laboratoires sont générées à partir de situations contrôlées et constituent souvent la base analytique typique de l'élaboration des modèles de synthèse.

Il n'est pas seulement nécessaire de mettre en évidence les éléments prosodiquement saillants, il faut également savoir *à quels endroits dans la parole* ces éléments deviennent importants. Encore une fois, nous devons être prudents dans la sélection des éléments de la variabilité que nous choisissons de modéliser et de ceux que nous nous permettons d'ignorer. (Certains chercheurs, par exemple, ont montré que des améliorations notables peuvent être obtenues dans la qualité de la synthèse si de petites variations micro-prosodiques, souvent ignorées, sont modélisées — voir exemple Monaghan, 1992.) L'une des premières expériences faites par Cutler (1976) montre par exemple qu'il faut être sélectif quant aux choix des éléments sur lesquels nous décidons de concentrer nos efforts de modélisation. Elle démontre de façon convaincante que les détails phonétiques présents dans une partie pré-tonique, apparemment sans intérêt, de contour de l'intonation permettaient aux auditeurs de prédire la structure prosodique globale de l'énoncé. Cutler avait enregistré des phrases telles que :

(1) she managed to remove the *dirt* from the rug but not the berry stains

et

(2) she managed to remove the dirt from the *rug* but not from their clothes

dans lesquelles le mot "dirt" reçoit un « accent haut » (proéminence de la hauteur tonique / nucléaire) dans le premier cas, et un « accent bas » (accent rythmique et absence de proéminence de hauteur) dans l'autre. Dans ces deux énoncés, les contrastes linguistiques sont réalisés par la présence de syllabes à proéminence de ton dans les mots en italiques ; la place de la syllabe à ton proéminent est conditionnée par la suite de l'énoncé.

Cutler a également enregistré une autre version des phrases stimulus qui était aussi « neutre » que possible avec un accent de niveau « intermédiaire » sur le mot-clé. Puis elle a coupé et collé cette version du mot clé dans (1) et (2). Dans le cadre d'une expérience, elle a demandé à des sujets de repérer la consonne /d/. Cutler a trouvé que les temps de réaction étaient sensiblement plus rapides lorsque le contour d'intonation pré-tonique était celui qui était associé à « l'accent haut ». Autrement dit, les temps de réaction plus rapides vont de pair avec les prédictions de poids sémantique. Cette « attente » dont les auditeurs de l'expérience de Cutler faisaient usage est clairement liée à la notion de « cohésion phonétique » que j'ai introduite ci-dessus. Des types particuliers de « colle » phonétique ou prosodique cimentent les différents éléments de la parole ensemble. Les corrélats acoustiques de l'information linguistique sont complexes, répartis sur des sections relativement longues du signal, et les auditeurs font apparemment usage de cette information pour créer des prédictions sur l'énoncé qui arrivera sous peu. Faire une prédiction correcte (ou fausse) à un endroit, peut avoir des implications importantes pour la bonne (ou mauvaise) perception de quelque chose à un autre endroit dans le message.

Même si nous laissons de côté les questions portant sur l'alignement temporel précis entre les contours prosodiques et les éléments syllabiques et segmentaux, ces exemples illustrent que certaines interactions phonétiques contribuent à donner à la parole son caractère naturel et sa cohésion acoustique. Il est bien connu qu'en anglais, par exemple, un segment placé immédiatement avant la frontière prosodique d'une locution aura tendance à être différent d'un segment équivalent placé après. Il aura naturellement tendance à être allongé, à avoir une amplitude plus basse (et à présenter un chute d'amplitude). De plus, il aura aussi tendance

à être produit avec une phonation différente. Dans un travail antérieur (Local & Kelly, 1986), nous avons montré de manière détaillée comment les auditeurs s'orientent vers ces détails au cours de leurs conversations continues.

De même, il apparaît que les segments du noyau seront longs dans les syllabes accentuées, mais que ces caractéristiques de durée se différencient de celles des segments placés avant une frontière (Van Santen, 1994). Ces segments présentent également des différences d'inclinaison spectrale, en conséquence de mouvements phonatoires différents (Stevens, 1995 ; Sluijter *et al.*, 1995) et résultant d'associations spécifiques avec différentes structures accentuées. Campbell et Black (1996) montrent que la variation prosodique a un effet significatif sur les caractéristiques spectrales de la parole et que l'on peut en tirer avantage dans la sélection d'unités pour la synthèse par concaténation. Autant d'éléments prosodiques à connaître et à respecter au cours de l'élaboration d'une synthèse plus naturelle de la parole !

6. Prosodies à domaine long, rythme et action à distance

J'ai déjà suggéré que la faiblesse de la parole synthétique est en grande partie liée à la variabilité avec laquelle les détails subtils spectro-temporels sont modélisés. Quand les êtres humains se parlent, il existe une relation étroite entre les mouvements de l'appareil phonatoire et les propriétés des sons qui en résultent. La parole naturelle est acoustiquement cohérente, en ce sens que ses fins détails spectro-temporels reflètent le fonctionnement de l'appareil phonatoire. Ce que nous avons tendance à classer comme « variabilité » ou comme « phénomènes phonétiques en distribution libre » — c'est-à-dire, la variation légitime et systématique d'un phénomène acoustique résultant d'actions articulatoires — ne correspond donc pas véritablement à un ensemble de phénomènes produits au hasard. Selon nous, cette variabilité ajoute des éléments d'information supplémentaires au signal.

Considérons par exemple la variation sub-phonémique des voyelles finales dans des mots comme : "whinny", "windy", "pally", "pantry", "city", et "seedy". On peut noter que les locuteurs de l'anglais britannique prononcent ces mots avec une différence systématique (Local, 1989). Les voyelles finales dans "windy" et "seedy" sont régulièrement produites avec une quantité plus fermée que celle qu'on retrouve dans "whinny" et "city". Certains locuteurs (ceux de Tyneside en Angleterre nord-centrale,

par exemple) les distinguent par deux monophtongues différentes — une voyelle antérieure fermée en opposition avec une qualité plus ouverte et postérieure — tandis que d'autres emploient des diphtongues plus fermées en opposition avec des plus ouvertes. Pourquoi en serait-il ainsi ? Tous ces mots sont dissyllabiques, accentués sur la première syllabe. Il ne s'agit pas seulement d'un effet d'harmonie avec la voyelle de la syllabe tonique, comme on peut le constater dans les exemples cités. Il s'agit plutôt d'un de ces petits détails phonétiques qui caractérisent la parole naturelle.

En dépit du fait que tous ces mots soient constitués de deux syllabes, avec accent sur la première, il faut remarquer que le rythme de ces paires est systématiquement différent. "Windy" et "seedy" possèdent ce qu'Abercrombie (1965) décrit comme des quantités rythmiques « égale-égale » alors que "whinny" et city" ont des quantités « brève-longue ». Comme Abercrombie le montre, cette variation rythmique peut être expliquée dans le cadre d'une analyse qui établit une distinction entre la structure des syllabes initiales de ces paires de mots. Celle de "windy", "seedy" et "Mary" est considérée comme « lourde » (voyelle brève + deux consonnes dans le premier cas, voyelle longue dans les deuxième et troisième cas), alors que celle de "whinny", "city" et "marry" est dite « légère » (voyelle brève et une seule consonne).

Ainsi nous pouvons conclure que la variation en voyelle finale est liée au poids de la syllabe précédente[3]. La variation rythmique dans ces paires de mots anglais ne provient pas seulement de certaines propriétés de durée de leurs constituants segmentaux (contrairement au cas du finlandais). Les différences spectrales et temporelles sont subtiles, mais si nous incorporons ces détails dans une synthèse, nous constatons dans des expériences de reconnaissance à choix ouvert ou de repérage de « segments » une augmentation notable des segments correctement identifiés. (Voir aussi les résultats dans (1) qui montrent que l'incorporation dans une syllabe d'une variation consonantique contextuelle, présentée en condition bruitée, peut accroître l'intelligibilité phonémique de la parole synthétique de 15% environ.)

Enfin considérons un dernier exemple de variation phonétique apparemment insignifiante — le cas du schwa dans la prononciation des énoncés suivants : "go to the picture" et "go to the park" (voir données en Local, en préparation). Dans ce cas, nous trou-

---

3. [33] présente les conséquences phonétiques d'une variation rythmique identique en finlandais.

vons des différences systématiques et perceptibles dans la qualité du schwa pour les mots "to the", cette qualité étant plutôt antérieure dans l'exemple avec "pictures" et plus centrale et postérieure dans l'exemple avec "park". Il apparaît que ces exemples sont conditionnés linguistiquement (i.e., qu'il n'y a donc pas de variation erratique). Au cours d'expériences où la séquence "to the" avait été extraite d'un énoncé, et collée dans l'autre, il ressort que des attentes plus importantes (un nombre significatif au niveau de 5%) se sont produites lorsqu'on a demandé à des sujets de repérer la voyelle accentuée dans les substantifs.

On peut généralement trouver des effets semblables avec le schwa. Par exemple dans la deuxième syllabe du mot "barber" dans les énoncés "the barber bought one" et "the barber beat one", les qualités des schwas semblent être conditionnées par le statut de la voyelle accentuée dans le mot qui suit ("bought/beat"). À nouveau, lorsqu'on incorpore ces effets dans la synthèse, on remarque une amélioration sensible par rapport au degré de naturel associé à ces énoncés, ainsi que dans l'identification de segments et de mots présentés en condition bruitée.

7. Conclusion

On pourrait soutenir que les choses dont je viens de parler sont des cas spéciaux ou des phénomènes très rares qui ne sont, de toute façon, pas intéressants pour ce qui concerne l'amélioration de la qualité de la parole synthétique. Cependant, je ne pense pas que ce soient des cas particulièrement spéciaux. L'analyse de la parole naturelle et spontanée dans les conversations continues montre que ces caractéristiques sont tout à fait fréquentes et qu'elles contribuent de manière systématique à la cohérence acoustique de la parole. Nos expériences à l'université de York ainsi que celles de nos collègues à Cambridge et à University College London montrent que lorsqu'on incorpore ces détails phonétiques, qu'on intègre la « colle » phonético-prosodique, et qu'on tient compte des relations interactives entre les différents niveaux linguistiques, la synthèse de la parole résonne beaucoup plus naturellement.

John LOCAL
Laboratoire de Phonétique Expérimentale
Département des Sciences du Langage
Université de York, G.B.
lang4@tower.york.ac.uk

Remerciements

Plusieurs idées (et certaines formulations) contenues dans cet article ont été directement inspirées par des conversations avec Sarah Hawkins, Jill House, John Kelly, Eric Keller et Richard Ogden. Je les remercie de leur créativité. Ils ne sont, bien sûr, aucunement responsables des bêtises qui pourraient s'y trouver.

Références

ABERCROMBIE, D. (1965). Syllable quantity and enclitics in English. In *Studies in Phonetics and Linguistics*. Oxford: Oxford University Press, 26-34.

ASSMAN, P.F., & SUMMERFIELD, Q. (1990). Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies. *JASA, 88,* 680-697.

BREGMAN, A.S. (1990). *Auditory Scene Analysis*. London: MIT Press.

BROWMAN, C.P., & Goldstein, L.M. (1989). Articulatory gestures as phonological units. *Phonology, 6,* 201-251.

CAMPBELL, N. & BLACK, N. (1996). Prosody and the selection of source units for concatenative synthesis. In J. Van Santen, R. Sproat, J. Olive & J. Hirschberg (eds), *Progress in Speech Synthesis*. New York: Springer-Verlag, 279-292.

COOKE, M., & BROWN, G.J. (1994). Separating simultaneous sound sources: Issues, challenges and models. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition*. Chichester: John Wiley & Sons, 297-312.

CUTLER, A. (1976). Phoneme monitoring reaction times as a function of preceding intonation contour. *Perception & Psychophysics, 20,* 55-60.

FIRTH, J.R. (1948). Sounds and Prosodies. (Reprinted in Palmer, 1970, 1-26.)

GIMSON, A.C. (1962). *An Introduction to the Pronunciation of English*. London: Arnold.

HAWKINS, S. & SLATER, A. (1994). Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech. *Proc. ICSLP 94, 1,* 57-60.

HAWKINS, S. (1995). Arguments for a nonsegmental view of speech perception. *Proc. ICPhS-95, 3,* Stockholm, 18-25.

KELLER, E., WERNER, S., ZELLNER, B., & KELLER, P. (1995). Description of the LAIP synthesis system. *Summary of activities (1991-1995) Laboratoire d'analyse informatique de la parole (LAIP),* 14. University of Lausanne.

KELLY, J. & LOCAL, J.K. (1986). Long-domain resonance patterns in English. *Proc. International Conference on Speech Input/Output, Institute of Electronic Engineers*, 304-308.

KELLY, J. & LOCAL, J.K. (1989). *Doing Phonology*. Manchester: Manchester University Press.

KOHLER, K. (1997). Parametric control of prosodic variables by symbolic input in TTS synthesis. In J. Van Santen, R. Sproat, J. Olive & J. Hirschberg (eds), *Progress in Speech Synthesis.* New York: Springer-Verlag, 459-475.

LOCAL, J. (1994). Phonological structure, parametric phonetic interpretation and natural-sounding synthesis. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition.* Chichester: John Wiley & Sons, 253-270.

LOCAL, J. (1996). Conversational phonetics: Some aspects of news receipts in everyday talk. In E. Couper-Kuhlen & M. Selting (eds), *Prosody in Conversation.* Cambridge: Cambridge University Press, 177-230.

LOCAL, J. (en préparation). On the linguistic relevance of sub-phonemic lawful variability: Data and analysis.

LOCAL, J., & KELLY, J. (1986). Projection and silences: Notes on phonetic and conversational structure. *Human Studies, 9,* 185-204.

LOCAL, J., & OGDEN, R. (1997). A timing model for non-segmental phonological structure. In J. Van Santen, R. Sproat, J. Olive & J. Hirschberg (eds), *Progress in Speech Synthesis.* New York: Springer-Verlag, 109-121.

LOCAL, J.K., & OGDEN, R. (à paraître). Nordic Prosodies: representation and phonetic interpretation. *Nordic Prosody VII.* Joensuu.

LOCAL, J.K., KELLY, J., & WELLS, W. (1986). Towards a phonology of conversation: Turn-taking in urban Tyneside speech. *Journal of Linguistics, 22,* 411-437.

LOCAL, J. (1989). Some rythm resonance and quality variations in urban Tyneside speech. In S. Ramsaran (ed.), *Studies in the Pronunciation of English: A commemorative volume in Honour of A.C. Gimson.* London: Croom Helm, 286-292.

MANUEL, S., SHATTICK-HUFNAGEL, S., HUFFMAN, M., STEVENS, K.N., CARLSON, R., & HUNNICUTT, S. (1995). Studies of vowel and consonant reduction. *Proc. ICSLP 92, 2,* 943-946.

MONAGHAN, A. (1992). Extracting microprosodic information from diphones — A simple way to model segmental effects on prosody for synthetic speech. *ICSLP 92,* 1159-1162.

SIMPSON, A. (1992). Casual speech rules and what the phonology of connected speech might really be like. *Linguistics, 30,* 535-548.

SLUIJTER, A., SHATTUCK-HUFNAGEL, S., STEVENS, K.N., & VAN HEUVEN, V. (1995). Supralaryngeal resonance and glottal pulse shape as correlates of stress and accent in English. *Proc. ICPhS-95, 2.* Stockholm, 630-633.

STEVENS, K.N. (1995). Prosodic influences on glottal waveform: preliminary data. *Speech Communication Group Working Papers.* RLE, MIT. vol. x.

SUMMERFIELD, Q., & CULLING, J.F. (1994). Auditory computations that separate speech from competing sounds. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition.* Chichester: John Wiley & Sons, 313-338.

VAN SANTEN, J. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language, 8,* 95-128.

VAN SANTEN, J., & Hirschberg, J. (1994). Segmental effect on timing and height of pitch contours. *ICSLP 94.*

WERNER, S., & KELLER, E. (1994). Prosodic aspects of speech. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition.* Chichester: John Wiley & Sons, 23-40.

ZELLNER, B. (1994). Pauses and the temporal structure of speech. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition.* Chichester: John Wiley & Sons, 23-40.