hc21

Synthèse de la parole à partir du texte

Synthèse de la parole à partir du texte

par **Christophe D'ALESSANDRO**

Directeur de recherches
LIMSI-CNRS, Orsay, France

et **Gaël RICHARD**

Professeur
Institut Mines-Télécom, Télécom ParisTech, CNRS-LTCI, Paris France

1.	**Histoire de la synthèse de la parole**	H 7 288v2 -	2
2.	**Texte – Analyses et traitements linguistiques**	—	3
2.1	Normalisation et prétraitement	—	3
2.2	Analyse lexicale et morpho-syntaxique	—	4
2.3	Analyse morpho-syntaxique	—	5
2.4	Analyse syntaxique	—	5
2.5	Transcription graphème-phonème	—	6
3.	**Signal de parole – Modèle source-filtre**	—	9
3.1	Modèle paramétrique de synthèse de parole	—	9
3.2	Caractéristiques du filtre	—	9
3.3	Caractéristiques de la source	—	10
4.	**Prosodie**	—	11
4.1	Prosodie et syntaxe	—	11
4.2	Calcul du rythme	—	12
4.3	Calcul de l'intonation	—	13
5.	**Synthèse acoustique**	—	13
5.1	Synthèse à formants par règles	—	14
5.2	Synthèse non paramétrique par concaténation d'unités acoustiques	—	15
5.3	Synthèse par diphones	—	15
5.4	Synthèse par sélection et concaténation	—	18
5.5	Synthèse paramétrique statistique	—	20
5.6	Construction du corpus textuel et sonore	—	22
6.	**Applications de la synthèse de parole**	—	23
6.1	Exemples d'applications	—	23
6.2	Interfaces de programmation	—	24
6.3	Produits	—	24
7.	**Évaluation de la synthèse**	—	25
7.1	Boîte noire ou boîte de verre	—	25
7.2	Évaluation de qualité globale	—	25
8.	**Conclusion**	—	26
8.1	Bilan	—	26
8.2	Perspectives	—	27
Pour en savoir plus		Doc. H 7 288v2	

L'objet de la synthèse de la parole à partir du texte (ou TTS, Text-To-Speech) est de calculer automatiquement le signal de parole correspondant à un texte donné. Le texte lui-même peut provenir de diverses sources : journaux, livres, systèmes de réponse vocale, de dialogue ou traduction automatique (borne interactive, assistant personnel), base de données d'un système d'information, jeu vidéo, courriers électroniques, SMS, documents butinés sur la toile, ou tout simplement texte saisi au clavier d'un ordinateur.

La réponse vocale sous sa forme la plus simple peut être un ensemble de messages préenregistrés (ou « prompts *»). L'ambition de la synthèse de la parole à partir du texte est plus grande : il s'agit de calculer automatiquement les échantillons sonores correspondant à un énoncé écrit quelconque, qui n'est pas connu d'avance et qui peut être de grande taille.*

Les deux versants de la synthèse de parole sont d'une part, l'analyse et l'interprétation du texte, d'autre part, la prédiction des paramètres acoustico-phonétiques du son et la synthèse du signal proprement dite :

– analyse du texte. *La première étape de la transformation d'un texte en parole implique la capacité d'analyser, de comprendre le texte écrit, ses nuances et ses connotations, la situation du discours et l'acte de parole à effectuer. En plus du texte, le contexte peut être spécifié (style de parole, émotion, attitude, type de personnage, voix spécifique...) ;*

– synthèse du signal. *Une fois le texte analysé, il s'agit de calculer le signal acoustique qui interprète au mieux le contenu linguistique, avec une voix aussi naturelle que possible, ressemblant à un locuteur particulier, et avec les nuances d'attitude, voire d'émotion que le texte réclame. En plus du signal audio, le synthétiseur peut fournir des indications pour synchroniser le mouvement des lèvres d'un avatar ou personnage vidéo, ou les mouvements d'un robot.*

1. Histoire de la synthèse de la parole

L'histoire de la synthèse à partir du texte est déjà longue : par exemple le premier système autonome de synthèse automatique de la parole en français, l'Icophone V du LIMSI, date de 1974. Cependant, cela reste encore un domaine de recherche très actif. Les travaux actuels portent à la fois sur la compréhension des textes et sur la restitution d'une parole naturelle, personnalisée et expressive. Les applications se multiplient. Le lecteur désireux d'approfondir les notions présentées dans cet article pourra consulter les références portées dans la partie documentation.

L'architecture générale d'un système de synthèse se compose ainsi de ces deux parties principales (figure **1**). Les principaux modules correspondant à ces traitements sont décrits dans cet article. Bien que tous les exemples cités dans la suite soient tirés du français, il est important de souligner que les problèmes posés sont similaires pour toutes les langues. Cependant, des différences notables existent en fonction des spécificités linguistiques de chaque langue particulière et notamment en ce qui concerne leur :

– notation graphique [alphabétique (romaine cyrillique, hébraïque, sanskrite, arabe...), syllabique (coréenne, japonaise...), idéogrammatique (chinoise, japonaise...)] ;

– grammaire (agglutinante, flexionnelle, niveaux de langue...) ;

– phonologie et phonétique (système de phonèmes, langues à ton, langues à clics...) ;

– prosodie (durées, intonation, qualité vocale).

Les systèmes actuellement développés pour toutes les langues s'inspirent de principes identiques, même s'ils diffèrent pour les corpus, lexiques, analyses et heuristiques linguistiques.

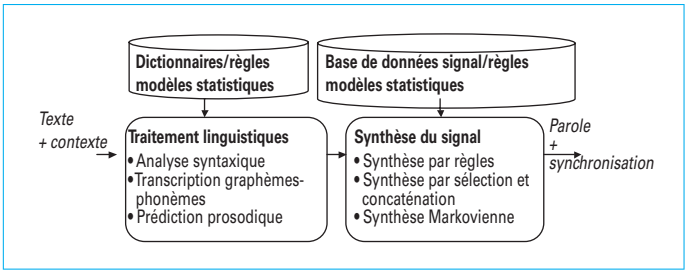


Figure 1 – Architecture générale d'un système de synthèse à partir du texte

2. Texte – Analyses et traitements linguistiques

La première étape d'un système TTS comprend les modules de traitements linguistiques qui permettent de transformer le texte à synthétiser en une chaîne de symboles représentant les sons distinctifs de la langue, ou « phonèmes », et un ensemble d'indications « prosodiques » caractérisant l'élocution (durée des différents sons et des pauses, évolution de la mélodie). Cette représentation phonético-prosodique est ensuite utilisée par l'étage de synthèse sonore qui assure la génération du signal de parole numérisé.

Les traitements linguistiques se décomposent en plusieurs étapes, qui obéissent en général à la séquence suivante :

– **normalisation et prétraitement du texte brut** : cherche les phrases, les anomalies du texte et produit une séquence lexicale (séquence d'unités lexicales ou « mots ») ;

– **analyse lexicale et morpho-syntaxique** : enrichit la séquence lexicale avec des étiquettes lexicales, des marques morpho-syntaxiques et produit un découpage en composantes syntaxiques ;

– **transcription graphème-phonème** : transforme la suite lexicale de sa forme orthographique à la forme phonétique.

2.1 Normalisation et prétraitement

Les sources du texte à synthétiser peuvent être variées : interfaces de programmation (API), textes issus de la toile, SMS, journaux ou livres électroniques, courriels, etc.

Confrontés à des sources de textes réelles et variées, le système de synthèse doit normaliser les nombreuses « anomalies » des textes au niveau graphémique (c'est-à-dire des « graphèmes » ou caractères écrits). Cette étape de prétraitement a pour objet de retranscrire en toutes lettres les chaînes de caractères non lexicales (hors dictionnaire), ou inconnues.

Il peut s'agir de nombres et de chiffres (3 999 356 à transcrire « trois millions neuf cent quatre vingt dix neuf mille trois cent cinquante six »), de dates (« 24/01/63 », « 24 janvier 1963 »), ou plus généralement de mots composés de caractères orthographiques et numériques (« vol AF102 », « référence SD44 »), de symboles spéciaux (« € », « © »).

Les diverses formes d'abréviations (par exemple, « c.-à-d. », « Pr. ») doivent aussi être repérées et traitées, comme les sigles (suite de lettres initiales non prononçable, qui est épelée, « CNRS », « SNCF »), les acronymes (suite de lettres initiales prononçable, « UNESCO », « ENA »).

L'analyse de la ponctuation est aussi un élément de prétraitement. L'étape de normalisation utilise à la fois un système de

règles de transcription (pour le traitement des quantités numériques ou des dates, des abréviations standards, ou des acronymes), et un lexique paramétré par l'utilisateur, spécifique de chaque domaine d'application de la synthèse.

Il est important de noter que beaucoup d'éléments sont ambigus. À titre d'exemple, il faut distinguer le rôle du point comme élément syntaxique (point final) ou comme symbole numérique (« 1.3 »). Seule une procédure de désambigüisation, utilisant des heuristiques contextuelles permet de décider de la transcription appropriée.

L'étape de prétraitement permet de former une suite de « lexèmes », ou unités lexicales à partir du texte d'entrée. Les signes de ponctuation et les caractères typographiques non littéraux sont également analysés, afin de former la suite lexicale (par exemple en séparant les lexèmes comme dans des formes composées comme « l'y » ou « 3,5 »).

L'apparition de textes électroniques peu structurés, voire entachés de fautes plus ou moins systématiques (courriers électroniques, textes sans accent, orthographe approximative, etc.) pose des problèmes nouveaux que l'on se doit de traiter automatiquement.

2.2 Analyse lexicale et morpho-syntaxique

Par le prétraitement, le texte d'entrée a été transformé en une suite d'unités lexicales, séparées, et encadrées par la ponctuation. L'étape d'analyse lexicale consiste à rechercher dans un lexique les informations associées aux différents lexèmes.

■ Le lexique contient les **classes lexicales** (ou parties du discours) associées au lexème. Suivant les systèmes, la classe lexicale peut être très simple, comme par exemple « mot outil/mot plein », qui sépare les mots qui n'ont qu'une fonction syntaxique de connexion et ceux qui ont une fonction sémantique (c'est-à-dire qui portent du sens). D'autres systèmes peuvent utiliser des catégories grammaticales « fines » qui combinent les grandes classes grammaticales (« nom », « adjectif », « verbe », « pronom »...) et des propriétés grammaticales de type (« genre », « nombre », « infinitif », « verbe d'état », « verbe transitif »...).

■ Il est fréquent que la catégorie grammaticale d'un lexème hors contexte soit ambiguë, c'est-à-dire que le même mot ait **plusieurs catégories grammaticales**.

Par exemple, un mot tel que « voile » peut être soit un nom masculin (« porter le voile »), un nom féminin (« larguer les voiles »), un verbe transitif à la première ou troisième personne de l'indicatif présent (« je voile », « il voile ») ou du subjonctif présent (« qu'il voile », « que je voile »), verbe pronominal (« le ciel se voile »).

Ainsi, plus de 70 % des lexèmes peuvent avoir plusieurs catégories grammaticales différentes.

■ La recherche lexicale est dans certains cas précédée d'une étape d'analyse morphologique, qui a pour objet de décomposer le lexème en composantes élémentaires, les **morphèmes**, correspondant aux **préfixes, suffixes, désinences** (marque du féminin ou du pluriel pour les noms et les adjectifs, temps, personne et mode pour les verbes), **racines**. On compte en français environ 500 préfixes, suffixes et désinences. Les désinences permettent généralement de déterminer des catégories grammaticales précises. Cela est particulièrement vrai pour les formes verbales conjuguées qui sont généralement composées d'une racine et d'indices grammaticaux, propres à la conjugaison du verbe, porteurs d'information sur le temps, le mode, et la personne. Dans les systèmes mettant en œuvre un traitement morphologique, le lexique se compose principalement de formes sources, ou radicaux (verbes, noms et adjectifs non fléchis, éventuellement privés de certains éléments de formation et/ou d'indices grammaticaux). Le lexique comporte aussi les exceptions de décomposition

morphologique, c'est-à-dire les mots qui ne se décomposent pas suivant les règles morphologiques du français standard.

La taille des lexiques varie significativement suivant les systèmes de synthèse. Certains systèmes utilisent des lexiques restreints, par exemple, aux mots outils et verbes (de quelques centaines à quelques milliers de mots). D'autres systèmes plus fins utilisent des lexiques de l'ordre de 50 000 à 100 000 mots (à titre de comparaison, un dictionnaire de langue comporte environ 50 000 articles). Les lexiques contenant toutes les formes fléchies, qui comprennent les mots et l'ensemble de leurs dérivations morphologiques se composent de 400 000 à 1 000 000 de mots et de locutions.

2.3 Analyse morpho-syntaxique

À l'issue du prétraitement des éléments non lexicaux et de la recherche lexicale, chaque lexème se trouve affecté à une ou plusieurs catégories grammaticales. Le choix de la catégorie de chaque mot s'effectue au moyen de règles contextuelles, ou de modèles statistiques, prenant en compte les catégories grammaticales des mots adjacents.

Les contextes sont réduits à quelques mots précédents ou suivants, on parle alors d'analyse morpho-syntaxique en micro-contextes. L'analyse des dépendances syntaxiques globales est beaucoup plus difficile à mettre en œuvre.

L'analyse syntaxique peut faire appel à des **règles heuristiques issues des règles de la grammaire de la langue** (par exemple « on ne peut observer la succession de deux verbes conjugués »).

Une autre stratégie consiste à suivre une approche statistique, exploitant des modèles probabilistes du langage. Ces modèles sont fondés sur l'observation que toutes les successions de catégories grammaticales dans une langue donnée ne sont pas équiprobables.

On peut donc chercher à résoudre les ambiguïtés en recherchant dans l'ensemble des successions possibles de catégories grammaticales (chaque mot est *a priori* porteur de plusieurs catégories possibles et l'on considère l'ensemble des transitions entre ces différentes catégories) la succession de catégories la plus probable.

Ces modèles probabilistes présentent l'avantage de ne requérir qu'une connaissance sommaire de la langue à traiter, à l'inverse des approches heuristiques qui compilent des connaissances et des observations très fines sur la structure des dépendances fonctionnelles des mots dans chaque langue. Cette caractéristique présente un avantage décisif, dès que l'on s'intéresse aux systèmes multilingues, les mêmes paradigmes d'apprentissage pouvant être déclinés pour plusieurs langues, sans remettre en question la structure du système.

2.4 Analyse syntaxique

L'analyse morpho-syntaxique permet la désambiguïsation des parties du discours, et associe à chaque unité lexicale une catégorie grammaticale.

Cette suite de catégories grammaticales permet de réaliser une analyse syntaxique de la phrase, c'est-à-dire de la découper en constituants syntaxiques et de grouper les mots.

Les groupes de mots permettent de structurer la phrase, en proposant un « parenthésage ». La constitution des groupes syntaxiques est effectuée par des règles heuristiques ou des modèles statistiques sur les successions de catégories pour former des groupes syntaxiques. Le groupement des mots est une étape essentielle pour le calcul de la prosodie (§ 4), des contours mélodiques et rythmiques de l'énoncé. Voici un **exemple** d'analyse syntaxique simple « en tronçon ».

Considérons la phrase :

Maintenant, un peu de voix féminine.

La suite de catégories grammaticales, ou parties du discours, associée est :

ADV PMK ART ADV ART NOM ADJ PMK

Avec : ART article, ADV adverbe NOM nom, ADJ adjectif, PMK marque de ponctuation. La liste des mot-outils et des mots pleins (nom, adjectif, verbe et adverbe) et les règles heuristiques de groupement permettent de regrouper les mots de la façon suivante.

(ART ADV ART ADV) (ART NOM ADJ) PMK

Grâce aux analyses lexicale, morpho-syntaxique, et syntaxique « en tronçons », le texte brut initial est maintenant structuré en sept mots avec une marque de ponctuation et deux groupes syntaxiques.

Le lecteur désireux d'approfondir ces notions consultera avec profit les références [1] [2].

2.5 Transcription graphème-phonème

Après la normalisation, étape qui transforme un texte brut en suite de mots sous la forme orthographique, et l'analyse lexicale et morpho-syntaxique, il s'agit de calculer la prononciation du texte. Cette « **orthographe inversée** » qui permet de passer des lettres aux sons a été abordée par le biais de systèmes de règles, de lexiques spécialisés, ou de techniques d'apprentissage automatique.

La transcription graphème-phonème, ou phonétisation, associe à la forme orthographique une chaîne de signes phonétiques qui spécifie la prononciation du mot.

Pour cela, est utilisé un « alphabet phonétique », sous-ensemble issu de l'alphabet phonétique international (API), et qui spécifie les sons élémentaires de la langue. Pour le français, un tel alphabet comporte 16 sons vocaliques, 20 sons consonantiques. Le codage informatique SAMPA est souvent utilisé en synthèse pour l'alphabet phonétique (tableau **1**).

■ Tout comme pour les catégories grammaticales, la même chaîne orthographique peut être associée à différentes transcriptions phonétiques : on parle alors d'« **homographes hétérophones** », mots qui s'orthographient de la même façon mais se prononcent différemment.

Le français standard comprend environ 150 homographes hétérophones ; il s'agit dans la plupart des cas d'ambiguïtés entre un verbe conjugué et un adjectif ou un adverbe formé sur la même racine, comme par exemple un président (nom)/ils président (verbe) ; somnolent (adjectif formé sur le verbe somnoler)/ils somnolent (verbe).

Le français comporte des cas, beaucoup plus rares, d'homophonie mettant en jeu des mots de racines différentes : les portions (nom)/nous portions (verbe) ; est (nom)/est (verbe).

Dans les cas précédents, la connaissance de la catégorie grammaticale du mot permet de choisir la prononciation correcte. Dans certains cas, plus rares, des homographes hétérophones ont la même catégorie grammaticale, comme par exemple : fils (du père)/fils (de coton). Seule une analyse sémantique (une étude du sens de la phrase) ou l'analyse du contexte élargi, permet alors d'effectuer la phonétisation correcte.

■ La **transcription phonétique** proprement dite est effectuée à l'aide d'un lexique d'exceptions et de règles. Après analyse des exceptions aux règles de phonétisation standards et des homographes hétérophones, les **règles générales de phonétisation** s'appliquent.

Un système de transcription orthographique-phonétique est un automate paramétré appliquant un ensemble de règles de réécriture,

Tableau 1 – Alphabet phonétique du français, avec les symboles issus du projet SAMPA

Description	Code phonétique SAMPA
archiphoneme / A /	A
[VANtardise, tEMPS]	A~
schwa	@
closed / oe / [crEUser, dEUx]	2
open / E / [pERdu, modEle]	E
closed / e / [Emu, otE]	e
[pEINture, matIN]	E~
[malhEUreux, pEUr]	9
[Idée, amI]	i
open / O / [Obstacle, cOrps]	O (o Majuscule)
closed / o / [AUditeur, bEAU]	O
[rONdeur, bON]	O~
[cOUpable, lOUp]	u
[pUnir]	y
[OUi, OIseau]	w
[hUIle]	H
[plétiner, paiLLe]	j
	p
	t
	k
	b
	d
	g
	f
	s
[CHanter, maCHine]	S
	v
	z
[Jardin, manGer]	Z
	l
uvular / R /	R
	m
	n
[aGNeau, règNe]	J
[campINg]	N

qui permettent d'associer un phonème (ou un groupe de phonèmes) à un caractère (ou un groupe de caractères) orthographique(s), en prenant en compte le contexte gauche (caractères ou groupes de caractères précédant le segment à transcrire) et le contexte droit

(caractères ou groupes de caractères suivant le segment à transcrire). Ces règles sont organisées de façon hiérarchique, des règles les plus particulières aux règles les plus générales.

Le nombre de règles nécessaires pour effectuer la transcription orthographique phonétique dépend de la langue considérée ; par exemple, moins de 100 règles sont requises pour une langue comme l'espagnol, la forme orthographique étant très proche de la forme phonétique.

Même si les exceptions ne sont pas prises en compte, le cas du français est beaucoup plus complexe, car la forme orthographique est, pour des raisons historiques, très éloignée de la forme phonétique. Un système minimal de description des règles de phonétisation du français standard se compose environ de 500 règles. Les meilleurs systèmes associent un lexique important avec de plus de 2 000 règles.

Pour donner un **exemple**, le mot « oiseau » se transcrit phonétiquement en /wazo/, par application des règles suivantes :

1. la chaîne de caractères orthographiques « oi » se transcrit par la succession des phonèmes /wa/, parce qu'elle est précédée d'un séparateur de mot et qu'elle n'est pas suivie de la chaîne « gn » comme dans « oignon », ou d'un « n » comme dans « oindre » ;

2. la lettre « s » se transcrit par le phonème /z/ car cette lettre est entourée par deux voyelles et que « oiseau » ne fait pas partie d'une liste d'exceptions à cette règle, stockée dans le lexique (comme « paraSol » ou « vraiSemblance ») ;

3. la chaîne de caractère « eau » se transcrit par le phonème /o/, indépendamment du contexte.

La chaîne phonétique obtenue après transcription peut être enrichie de marques syllabiques, à l'aide d'un petit ensemble de règles. La syllabe est souvent considérée comme l'unité rythmique de base, utile pour le calcul de la prosodie.

La question de la **phonétisation**, qui semble en apparence assez régulière, se révèle en fait difficile lorsqu'il faut traiter de problèmes comme les noms propres (notamment ceux d'origine étrangère), les mots nouveaux et inconnus, les variantes de prononciation, les différents dialectes, idiolectes ou sociolectes. Cela pose d'importantes questions de phonologie, comme celles du « e » muet, de la coupe syllabique, des liaisons, de l'harmonie vocalique, de l'emprunt de phonèmes d'autres langues, etc. Même avec des taux de phonétisation correcte très élevés (plus de 95 ou 98 %), les erreurs effectives de phonétisation restent fréquentes dans un texte de quelques secondes, puisque nous prononçons environ 10 phonèmes par seconde.

Voici un **exemple** de chaînes obtenues à la **sortie du module de phonétisation** :

La légende veut que, en rêvant devant Grenade au cours de vacances espagnoles, lord Sydney Bernstein ait choisi le nom de son entreprise.

Texte :

Cinquante ans plus tard, le groupe pèse un peu plus de 8 milliards de francs et affiche un bénéfice de 900 millions pour 1986.

Phonèmes et mots :

sIkAt A ply tar, lx grup pEz Ip@ ply dx hi miljar dx frA e afiS I benefis dx nXf sA miljO pur mil nXf sA katrx vI sis.

Catégories grammaticales des mots :

ADJ| NOM| ADV| ADV| VRG| PPL| VCO| VCO| ADV| ADV| PRP| ADJ| NOM| PRP| NOM| CCO| VCO| DET| NOM| PRP| ADJ| ADJ| NOM| PRP| ADJ| ADJ| ADJ| ADJ| ADJ| ADJ| PMK|

3. Signal de parole – Modèle source-filtre

Cette approche est fondée sur un modèle source/filtre de production du signal vocal, **modèle commandé par un nombre restreint de paramètres**. La synthèse se décompose en deux étapes :

1. à l'aide de règles contextuelles, les informations phonético-prosodiques sont transformées en commandes permettant de spécifier l'évolution temporelle des paramètres du modèle de synthèse ;
2. les valeurs des paramètres ainsi déterminées sont utilisées par le vocodeur pour synthétiser le signal acoustique.

Historiquement, dès la fin des années 1950, ce type de technique a été la première à émerger avec les vocodeurs de l'époque, et des règles élaborées explicitement par l'analyse de corpus phonético-acoustique. Elle reste encore utilisée, après une éclipse en faveur des techniques par concaténation d'unités (§ 4), grâce aux techniques statistiques qui remplacent avantageusement les règles explicites par des techniques d'apprentissage automatique.

3.1 Modèle paramétrique de synthèse de parole

Le signal de parole peut être modélisé par un système source-filtre. La parole résulte de l'excitation des **cavités acoustiques supra-glottiques** (conduit oral, conduit nasal) par des impulsions acoustiques créées par le flux d'air en provenance des poumons et modulé par les cordes vocales. Le modèle source-filtre de la parole représente la production vocale en distinguant deux éléments : une source de phonation, qui représente les impulsions ou le bruit généré à la glotte, et un filtre acoustique qui modélise la contribution de la partie articulatoire.

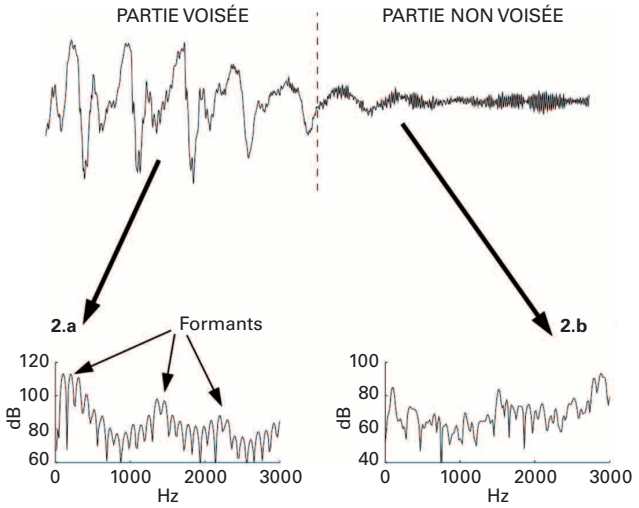
3.2 Caractéristiques du filtre

Les cavités supra-glottiques jouent le rôle d'un **résonateur acoustique**. Les caractéristiques acoustiques de ce résonateur (les fréquences amplifiées et atténuées par ce dispositif) dépendent de la **géométrie du conduit oral et de son degré de couplage avec le conduit nasal**. Ces caractéristiques géométriques sont elles-mêmes contrôlées par les **articulateurs**, à savoir les lèvres, la langue, et la mâchoire. Le couplage entre la cavité orale et la cavité nasale est contrôlé par l'intermédiaire du voile du palais (ou **velum**), qui peut s'abaisser ou se relever, réglant ainsi le débit du flux d'air dévié dans les fosses nasales.

Les caractéristiques acoustiques des cavités supra-glottiques peuvent être, en première approximation, modélisées à l'aide d'un filtre linéaire dont la fonction de transfert varie au cours du temps. Les variations de la fonction de transfert reflètent les modifications de la géométrie des cavités supra-glottiques au cours de l'articulation des différents phonèmes constituant l'énoncé.

Du fait de l'inertie mécanique de ces articulateurs et de la nature de leur contrôle musculaire, les variations sont relativement lentes : le temps typique de stabilité des caractéristiques spectrales est la dizaine de millisecondes (10 millisecondes pour les évènements les plus brefs, 40-50 millisecondes pour les segments les plus stables). En pratique, il suffit de spécifier ces fonctions de transfert toutes les 10 à 20 millisecondes.

Les paramètres les plus fréquemment utilisés pour contrôler les caractéristiques de ce filtre sont ceux des « formants » spectraux, ou maxima de la fonction de transfert du conduit vocal, à savoir la fréquence centrale, la bande passante et l'amplitude des maxima (figure **2**). Pour obtenir une parole intelligible, il suffit de spécifier les valeurs des trois à quatre premiers formants.



Notons que la forme d'onde est quasi périodique dans la partie voisée (à gauche du trait pointillé). 2 Estimations de la densité spectrale de puissance du signal obtenues par transformée de Fourier (entre 0 et 3 ~ kHz), pour la partie voisée (2.a) et pour la partie non-voisée (2.b). Pour la partie voisée (2.a), on observe les harmoniques du signal ainsi que la structure des formants (désignés par les flèches).

Figure 2 – Forme d'onde d'un signal de parole (environ 100 ms)

3.3 Caractéristiques de la source

L'excitation du conduit vocal, considéré comme un filtre acoustique avec plusieurs résonances, dépend du type de phonème considéré (voisé ou non voisé), du mode de phonation (chuchoté, crié...) et de l'effort vocal.

■ Pour les **phonèmes voisés**, **les voyelles et les consonnes voisées**, les plis vocaux, un peu comme les lèvres du trompettiste, modulent de façon périodique le débit d'air s'écoulant à travers la glotte. La fréquence de cette modulation est la fréquence de voisement, ou fréquence fondamentale. Pour des raisons aérodynamiques, cette onde de débit glottique est généralement très dissymétrique, comprenant une phase (dite d'**ouverture**) où le débit augmente lentement, suivie d'une phase (dite de **fermeture**) plus abrupte. La fréquence de vibration, ainsi que la forme de l'onde de débit est contrôlée par les muscles et cartilages dont est constitué le larynx. L'onde de débit glottique est représentée à l'aide d'un modèle paramétrique contrôlé par la période de voisement, l'amplitude, et des paramètres de forme.

■ Pour les **phonèmes non voisés, les consonnes non voisées**, les cordes vocales restent ouvertes pour permettre le passage d'un flux d'air en provenance des poumons ; l'excitation est due, soit au relâchement rapide d'une occlusion complète du conduit vocal (**plosive**), soit aux turbulences du flux d'air créées au passage d'une constriction du conduit vocal (**fricatives**). Ces signaux sont modélisés par des sources de bruit, transitoires ou continues, réparties dans le conduit vocal, sources de bruit dont on contrôle à la fois la position et la puissance.

■ Notons finalement que, dans la plupart des segments mais plus particulièrement dans les fricatives et plosives voisées (en français /b/ /d/ /g/, /v/ /z/ /j/), ces deux modes d'excitation (voisée/non

voisée) coexistent. On synthétise alors le signal d'excitation en combinant l'onde de débit glottique et les sources de bruit.

4. Prosodie

La prosodie est la **« musique » de la parole**, c'est-à-dire sa composante mélodique, rythmique et dynamique. Une même chaîne de phonèmes peut être prononcée sur des « tons » très différents, par la variation prosodique. Du point de vue de la synthèse, le calcul prosodique consiste à modéliser et prédire :

– les contours mélodiques, par l'évolution temporelle de la fréquence fondamentale de vibration des cordes vocales ;

– le rythme syllabique, par les durées des syllabes et des phonèmes ;

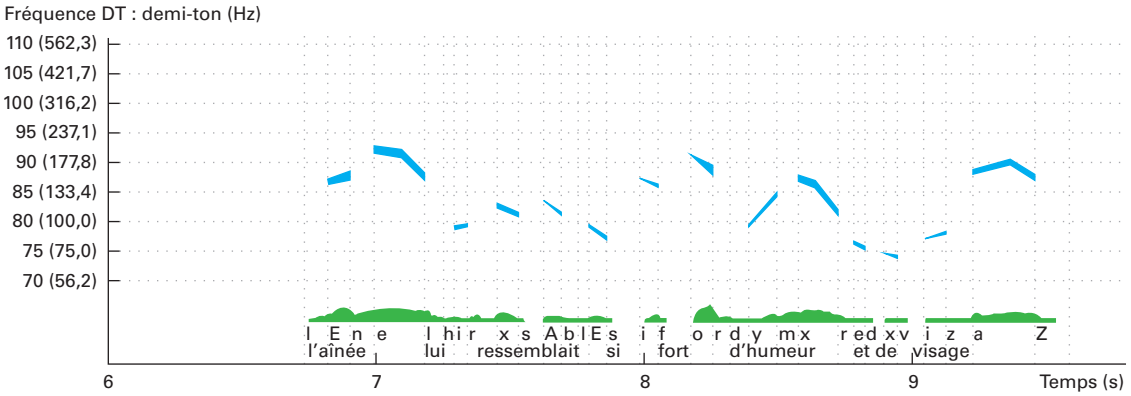
– le rythme des groupes de mots, par les positions et les durées des pauses ;

– la dynamique, l'effort vocal, l'intensité relative des syllabes ;

– la qualité vocale incluant le murmure, le chuchotement, les différents mécanismes de voisement ;

– les éléments extralinguistiques, comme les soupirs, rires, bruits de bouche, respiration, raclement de gorge, pauses vocalisées (« hummm »).

La prosodie joue un rôle important pour le naturel de la voix, mais aussi pour son intelligibilité. Certains systèmes de synthèse actuels offrent des styles vocaux différents, des voix dynamiques, enjouées, ou au contraire accueillantes, de proximité. Ces effets sont essentiellement des effets prosodiques. La modélisation prosodique est donc une **composante tout à fait essentielle d'un système de synthèse de parole** (figure **3**).



En haut, contours mélodiques stylisés de chaque syllabe. Milieu : courbe d'énergie. Texte : transcription phonétique alignée sur le signal, et transcription orthographique (« L'aînée lui ressemblait si fort d'humeur et de visage », voix masculine).

Figure 3 – Prosodie d'une phrase

4.1 Prosodie et syntaxe

Tout comme la phrase syntaxique s'ordonne de façon hiérarchique en groupes syntaxiques (groupe sujet, groupe verbal, groupe complément), la phrase prosodique s'organise en groupes prosodiques, groupes de mots le plus souvent séparés par des pauses.

Pour chaque groupe, les paramètres prosodiques (à savoir intonation, durées, dynamique, qualité vocale) suivent une évolution particulière, dépendant du rôle du groupe prosodique dans la

phrase, de sa position et de ses dépendances fonctionnelles avec les groupes adjacents, du nombre de syllabes, mais aussi du sens de la phrase, du style et de l'intention du locuteur.

Les frontières des groupes prosodiques ne reflètent pas systématiquement les frontières des groupes syntaxiques. Une certaine congruence peut toutefois être notée, surtout pour ce qui concerne les frontières syntaxiques majeures (frontière de phrase ou de clause, mais aussi frontière entre le groupe sujet et le groupe verbal associé).

■ La **structuration prosodique de la phrase** obéit à des règles moins bien formalisées que la structuration syntaxique, puisqu'elle ne dépend pas uniquement de la syntaxe, mais aussi du sens (sémantique linguistique), de l'interaction (pragmatique linguistique), du style de parole. Alors qu'il n'existe, une fois les règles grammaticales définies, qu'une seule façon d'analyser syntaxiquement une phrase, il existe plusieurs façons de décomposer la phrase en groupes prosodiques et plusieurs façons de structurer l'évolution des paramètres prosodiques sur chacun de ces groupes.

Par exemple, le sens du message dont est porteur la phrase et l'intention du locuteur (le ou les points clefs de la phrase sur lesquels le locuteur cherche à donner de l'emphase) jouent un rôle tout aussi important que la structure syntaxique pour calculer la prosodie.

■ Il est actuellement difficile, sauf dans des domaines restreints, d'analyser automatiquement les aspects sémantiques ou pragmatiques des énoncés. Pour cette raison, le calcul de la prosodie s'appuie sur le **découpage de la phrase en groupes prosodiques en fonction du découpage syntaxique**. Les contours intonatifs et rythmiques s'appuient sur la syntaxe et la ponctuation. Ce type de prosodie semble convenable pour la lecture de textes, de messages (consultation vocale de messagerie, par exemple) ou d'informations générales (journal téléphoné, bulletin météo...). La prosodie obtenue est néanmoins vite lassante, et peu sensible au contexte d'élocution.

4.2 Calcul du rythme

Les pauses sont les temps de silence, de durée variable (de 100 ms à plusieurs secondes), qui s'insèrent à la fin de chacun des groupes prosodiques. L'importance de la coupure syntaxique liée à un marqueur prosodique détermine la durée de la pause à insérer. Ce facteur est particulièrement important pour le **naturel de l'élocution**.

■ Pour le rythme, en plus des pauses, est associée à chaque segment (phonème, syllabe) une durée. Cette durée est déterminée en prenant en compte différents facteurs, en particulier, la **durée intrinsèque des sons** constituant le segment et le contexte. La durée intrinsèque correspond à la durée moyenne du segment tout contexte confondu (ou dans un contexte neutre). Cette durée intrinsèque est généralement déterminée en analysant un **corpus prosodique**, c'est-à-dire un ensemble de phrases qui ont été segmentées et annotées. En plus de la durée intrinsèque, le contexte influence la durée d'un phonème : il peut s'agir de la nature des phonèmes adjacents (certains phonèmes ont tendance à allonger les phonèmes adjacents, d'autres à les raccourcir), de la position de la syllabe porteuse dans le groupe prosodique (en français par exemple, la syllabe finale des mots est généralement allongée, d'un facteur d'autant plus important que le groupe précède une frontière syntaxique majeure), de la nature du groupe prosodique (sa fonction dans la phrase), de la longueur du groupe prosodique, etc.

Les éléments de contexte sont souvent combinés par des « sommes de produits », qui permettent de prendre en compte les allongements ou raccourcissements contextuels. Par exemple, la durée d'une voyelle V, suivie d'une consonne C dans une syllabe finale de phrase sera calculée par :

Durée (V, contexte droit C, syllabe finale) = (durée intrinsèque V) + (consonne C) × (contexte droit) + [allongement final × (consonne C et voyelle V]

■ Une bonne détermination des durées segmentales est cruciale pour assurer le naturel de l'élocution (des durées erronées produisent une parole heurtée, chaotique et, parfois, difficilement intelligible).

Les durées des phonèmes et des syllabes peuvent être calculées par des systèmes de règles, par des tables de durée, ou par des méthodes d'apprentissage, souvent des arbres de classification et de régression (CART : *Classification And Regression Trees*).

4.3 Calcul de l'intonation

Le calcul de l'intonation, ou contour mélodique, est particulièrement important pour la **qualité de la synthèse**. L'intonation est la variation de fréquence de voisement, la fréquence de vibration des plis vocaux. Plusieurs modèles ont été proposés pour décrire cette fonction continue du temps en termes discrets, tout comme le flux continu de la parole est décrit par une chaîne de phonèmes :

– pour l'anglais et d'autres langues, le système ToBI (*Tones and Boundary Indexes*) discrétise la courbe intonative, en la représentant par un petit ensemble de tons hauts, bas, accentués (To), et des niveaux de frontière (BI) ;

– le modèle tonal perceptif ou Prosogram est un système de notation basé sur un modèle de perception de la hauteur tonale, prenant comme unité de base la syllabe. Les courbes mélodiques sont réduites par stylisation, et des tons statiques ou dynamiques obtenus sont affectés à chaque syllabe ;

– INTSINT (*INternational Transcription System for INTonation*) est un système de notation intonatif à visée multilingue, basé sur un système de stylisation mélodique par points cibles (MoMel).

Tout comme les durées, l'**évolution de la fréquence fondamentale** pour chaque phonème dépend de facteurs, reflétant le contexte local et global. On peut donc la calculer par règles ou par des méthodes d'apprentissage. Au niveau local, l'évolution de la fréquence fondamentale est essentiellement influencée par la nature du phonème, par sa position dans la syllabe et par son environnement phonétique immédiat (certains phonèmes, comme les occlusives voisées, contribuent à abaisser la fréquence fondamentale ; d'autres ont tendance à l'augmenter). Au niveau global (groupe prosodique, phrase), les facteurs importants sont la position de la syllabe dans le groupe prosodique (en français, la première et la dernière syllabe d'un groupe prosodique obéissent à des règles spécifiques ; dans les langues à accent fixe, les syllabes accentuées sont définies pour chaque mot dans le dictionnaire), la fonction du groupe prosodique dans la phrase, et le mode de la phrase (interrogatif, déclaratif...).

Il est important de noter que la modélisation prosodique est étroitement liée à la technique de synthèse acoustique utilisée. Pour la synthèse par règles ou par concaténation de diphones, la prosodie est en général calculée par règle. Pour la sélection et concaténation d'unités non uniformes, la prosodie est obtenue sans calcul explicite, puisque la sélection puis concaténation des unités choisies dans un corpus de grande taille conserve la prosodie originale des segments sélectionnés.

Dans la synthèse par modèles statistiques, la prosodie est calculée par apprentissage sur un gros corpus, et générée par exemple, à l'aide des chaînes de Markov.

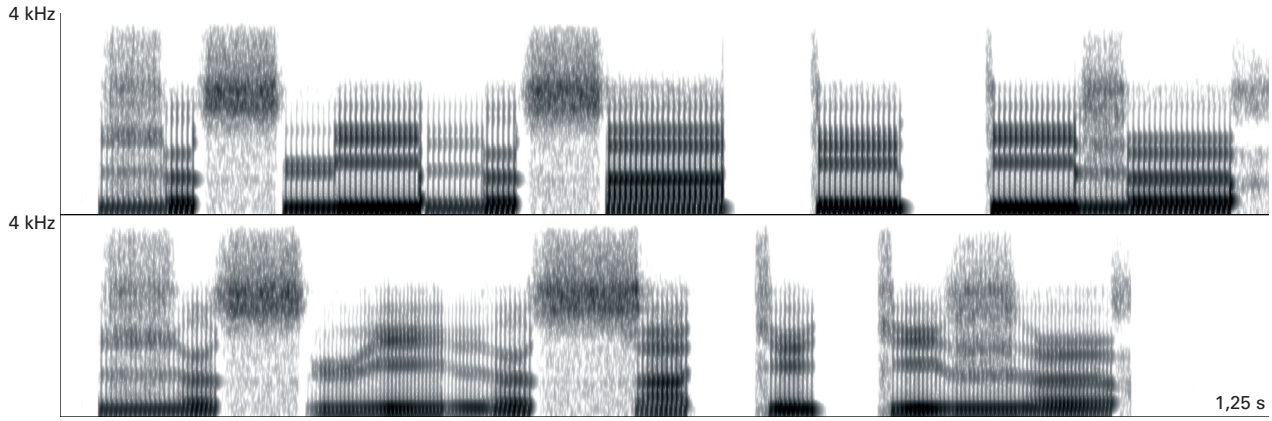
5. Synthèse acoustique

Après la phase d'analyse du texte à produire, intervient la phase de synthèse acoustique, qui consiste à **transformer la suite de**

symboles obtenus lors de l'analyse linguistique en suite d'échantillons de signal. Deux grandes classes de techniques existent pour la synthèse acoustique : la synthèse paramétrique et la synthèse non paramétrique. En synthèse paramétrique, le signal est calculé en utilisant un modèle du signal de parole, le modèle source-filtre : le signal passe par un « vocodeur » (pour « *voice coder* », ou système d'analyse-synthèse). En synthèse non paramétrique, les échantillons de signal sont concaténés avec des modifications minimes, sans passer par un vocodeur.

5.1 Synthèse à formants par règles

Le synthétiseur par règles calcule l'évolution (les trajectoires) des paramètres de contrôle du modèle de production à partir de la représentation phonético-prosodique, qui spécifie la chaîne des sons à prononcer, leur durée et le contour mélodique. La stratégie généralement utilisée consiste à spécifier tout d'abord des points cibles sur les segments stables du signal de parole (par exemple, valeurs de la fréquence centrale, de la bande passante et de l'amplitude de chaque formant au centre des voyelles) (figure **4**).



Spectrogramme des signaux de synthèse pour la phrase « Je suis le synthétiseur » en utilisant les valeurs cibles seules (haut), en utilisant les valeurs cibles, les règles de coarticulation, et les règles prosodiques (bas).

Figure 4 – Synthèse à formants par règles

Des règles d'interpolation des paramètres entre les différents points cibles sont ensuite mises en œuvre, modélisant les phénomènes de coarticulation, c'est-à-dire les interactions acoustiques entre phonèmes adjacents. Ces phénomènes de coarticulation sont la traduction acoustique des contraintes articulatoires, c'est-à-dire de la dynamique des articulateurs, le conduit vocal évoluant relativement lentement.

La synthèse proprement dite est réalisée à l'aide d'un **synthétiseur à formants** qui comprend :

1. un **module source**, qui comprend un générateur d'impulsions, pour la parole voisée, et un générateur de bruit blanc gaussien pour la parole non voisée. Les impulsions sont mises en forme par un modèle de l'onde de débit glottique (ou alternativement d'un modèle de la dérivée de l'onde de débit glottique), fonctions construites à l'aide de sinusoïdes, d'exponentielles ou de polynômes par morceaux. Bruits et impulsions peuvent se mélanger ;

2. un **module de filtre**, qui comprend en général cinq ou six résonateurs du second ordre, combinés en série ou en parallèle,

pilotés par les valeurs des formants, calculées par les règles de synthèse.

■ **La qualité obtenue en synthèse par règles est limitée**. D'une part, la mise au point de règles de synthèse performantes est une tâche longue et experte : l'obtention de résultats convaincants nécessite plusieurs années d'efforts pour un expert phonéticien, un handicap certain lorsque l'on cherche à mettre en œuvre des systèmes multilingues ou multi locuteurs. D'autre part, malgré de nombreux efforts en ce sens, aucun système de règle n'est suffisamment complexe et précis pour obtenir une qualité de voix complètement naturelle. Ainsi, dans le meilleur des cas, cette parole peut être très intelligible, mais assez peu naturelle.

Les lecteurs désirant approfondir les aspects liés aux modèles ou aux stratégies de commande les plus couramment utilisés dans le cadre de la synthèse par règles peuvent se reporter à la référence [3].

5.2 Synthèse non paramétrique par concaténation d'unités acoustiques

L'idée de restituer de la parole enregistrée en recombinant des échantillons est apparue dès les premiers systèmes de réponse vocale. L'avantage est la qualité naturelle de la parole obtenue, puisque ce sont des échantillons. Les inconvénients sont l'espace mémoire requis (ce qui n'est plus vraiment un problème aujourd'hui), et la nécessité d'enregistrer un même locuteur si on veut compléter la base de données.

Cette seconde approche n'utilise pas de modèle de production de la parole et ne dépend pas de paramètres d'un synthétiseur. Elle consiste à synthétiser le signal par concaténation d'unités acoustiques, c'est-à-dire de **segments de parole préenregistrés**. Cette technique, reposant sur l'utilisation de segments de signaux extraits de la parole naturelle, est la seule qui permette à ce jour de synthétiser des voix dont le timbre est proche, voire identique à celui d'un locuteur humain.

Les premiers systèmes de réponse vocale de haute qualité utilisaient de la « **synthèse à parties manquantes** », c'est-à-dire des phrases porteuses, dont certaines parties seulement étaient variables. Certaines applications, comme lire des bulletins météo ou bien des itinéraires, fonctionnent très bien de cette façon. Cependant, il ne s'agit pas vraiment de synthèse à partir du texte, puisque le vocabulaire est forcement limité.

5.3 Synthèse par diphones

Pour la synthèse par concaténation, il faut donc utiliser des unités plus petites que le mot. L'analyse linguistique a montré que le phonème est l'unité minimale de base de la parole. Mais ces unités, en petit nombre, sont en fait trop courtes et inappropriées car elles ne permettent pas de capturer la dynamique du processus de production de parole : la parole est **essentiellement un processus temporel continu** et la coarticulation entre sons voisins joue un rôle fondamental.

■ Ainsi, l'**unité minimale permettant d'obtenir une synthèse de qualité acceptable est le « diphone »**, qui est défini comme la portion du signal de parole comprise entre les noyaux stables de deux phonèmes consécutifs. Le diphone, à l'inverse du phonème, capture la transition entre les différentes cibles articulatoires associées aux phonèmes, transitions qui sont cruciales pour la perception des différents sons. En théorie, le nombre de diphones est égal au carré du nombre de phonèmes, c'est-à-dire environ à 1 300

(36 × 36) pour le français (en sachant que certaines transitions entre phonèmes sont en fait impossibles en français).

En pratique, pour la synthèse par diphones, le nombre d'unités utilisées est légèrement plus important (de l'ordre de 1 500-2 000) pour tenir compte des différentes variantes contextuelles des phonèmes composant le diphone (dans certains systèmes, comme plusieurs représentants de chaque diphone sont disponibles, l'algorithme de concaténation choisi à chaque instant le « meilleur » représentant de façon à minimiser une fonction d'objectif). Le volume de stockage nécessaire est de l'ordre de 5 à 10 Mo (2 à 6 min de parole numérisée avec une fréquence d'échantillonnage de 16 kHz). Cette quantité de données, qui a longtemps été considérable par rapport aux tailles mémoire disponibles, est petite en regard des possibilités de stockage offertes par les systèmes informatiques actuels.

■ Pour accroître la qualité, il est possible de considérer des **unités plus longues que le diphone**, aptes à prendre en compte des phénomènes de coarticulation à plus long terme (disons, pour simplifier, à l'échelle de la syllabe). Parmi celles-ci, les unités de la forme voyelle-consonne-voyelle (V-C-V), ou de façon plus générale du type V-C-... -C-V (deux voyelles séparées par un nombre quelconque de consonnes) permettent de n'avoir à effectuer des concaténations que dans les zones les plus stables du signal de parole, à savoir le centre des noyaux vocaliques. Elles capturent d'autre part la coarticulation de voyelle à voyelle à travers la (ou les) consonne(s), coarticulation qui joue un rôle important à la fois pour l'intelligibilité et l'agrément de la voix de synthèse. Le problème est que le nombre d'unités ainsi obtenues est beaucoup plus important (de l'ordre de 10 000-15 000, en ne retenant que les unités apparaissant effectivement). Un grand nombre de ces unités sont peu fréquentes et peuvent être éliminées pour satisfaire aux contraintes de taille.

■ La constitution du dictionnaire d'unités acoustiques se fait en enregistrant un **corpus de logatomes** (successions élémentaires de sons de parole n'ayant pas nécessairement de signification), qui servent de contexte aux unités choisies. L'extraction des unités acoustiques nécessite la segmentation des logatomes enregistrés. Celle-ci se fait de manière automatique en alignant, à l'aide de méthodes statistiques dérivées de la reconnaissance de parole, la transcription phonétique du mot et la forme acoustique. L'édition manuelle des résultats de segmentation, longue et fastidieuse, est nécessaire pour obtenir une synthèse de bonne qualité.

La synthèse proprement dite comprend trois étapes distinctes :

1. **Sélection des unités acoustiques**

Cette première étape consiste à choisir dans le répertoire d'unités acoustiques les unités qui seront effectivement utilisées pour synthétiser la succession de sons désirée. À partir de la représentation phonétique de l'énoncé, il s'agit de rechercher la suite de segments correspondants. Si les diphones sont utilisés, seule la présence de plusieurs versions pour le même segment est à prendre en considération. Cette étape est en revanche plus délicate pour les systèmes à base d'unités de taille variable. Pour une suite de sons donnée, plusieurs choix d'unités sont en général possibles. Il faut alors arbitrer entre les différentes décompositions avec des critères composites.

2. **Ajustement des paramètres prosodiques**

Les unités acoustiques préenregistrées possèdent une prosodie intrinsèque. Cette prosodie intrinsèque doit être neutralisée, afin d'appliquer la prosodie de synthèse spécifiée par le module prosodique. Dans ce cas, il est nécessaire d'utiliser une technique de traitement de signal pour ajuster aux valeurs cibles définies les paramètres prosodiques des unités de synthèse. Un tel système est décrit dans le paragraphe suivant.

3. **Concaténation des unités**

Les unités acoustiques, quelles que soient les précautions prises lors de la sélection et de l'enregistrement des unités, ne possèdent

pas exactement à leur frontière les mêmes caractéristiques acoustiques (en particulier énergétiques). En l'absence de traitement, ces discontinuités vont engendrer des artefacts perceptibles et gênants. Il est donc important de lisser ces discontinuités au moment de la concaténation.

■ Dans un système de synthèse par diphones, les variations prosodiques déterminées lors de la phase d'analyse linguistique doivent être appliquées aux unités acoustiques. Un système de **synthèse par concaténation** (par opposition à la synthèse par règles) n'implique pas l'utilisation d'un modèle de production du signal de parole. Les modifications de durée et de fréquence fondamentale du signal utilisent des techniques non paramétriques.

Une des premières techniques non paramétriques de modification prosodique est l'algorithme PSOLA (pour *Pitch-Synchronous Overlapp-Add*). La caractéristique la plus remarquable de la méthode PSOLA est qu'elle opère directement sur la forme d'onde du signal de parole. L'idée de base est d'extraire du signal des grains de sons élémentaires, représentant les caractéristiques locales du signal, et de déplacer ces grains de sons pour réaliser les modifications désirées. De façon plus précise, l'algorithme procède en trois étapes :

1. **Analyse**

Cette étape consiste à extraire du signal une suite de grains de sons élémentaires (signaux à court terme). Ces grains de sons sont obtenus en multipliant le signal par une fenêtre d'analyse centrée autour d'instants d'analyse. Les instants d'analyse sont disposés de façon synchrone à la fréquence fondamentale dans les segments de parole voisés. Ils sont répartis de façon non uniforme et arbitraire sur les segments non voisés.

2. **Transformation**

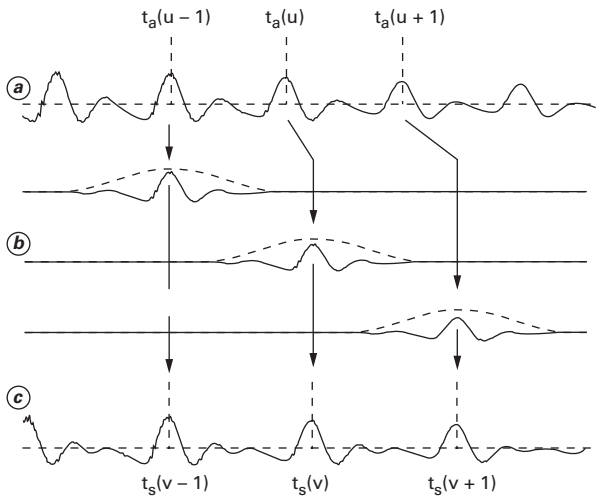
Cette étape consiste à calculer une suite d'instants de synthèse et une fonction d'association des instants de synthèse et des instants d'analyse pour réaliser la conversion désirée de durée et de fréquence fondamentale. On synchronise ensuite les signaux à court terme d'analyse sur les instants de synthèse, en utilisant la fonction d'association. On définit ainsi une suite de signaux à court terme de synthèse synchronisés sur les instants de synthèse. En l'absence de modification, les instants de synthèse correspondent aux instants d'analyse, et les signaux à court terme de synthèse sont égaux aux signaux à court terme d'analyse. La figure **5** illustre les caractéristiques de la fonction d'association dans le cas d'une modification simple de la fréquence fondamentale (abaissement par un facteur constant).

3. **Synthèse**

Cette dernière étape consiste à recombiner les signaux à court terme de synthèse. On procède en additionnant les échantillons des signaux à court terme de synthèse associés au même instant de synthèse. Le facteur de normalisation variable tient compte des variations d'énergie liées à la cadence irrégulière de l'analyse et de la synthèse. On remarque qu'en l'absence de modification, le signal de synthèse correspond exactement au signal d'analyse.

■ Le **coût de calcul associé à la méthode PSOLA** est très raisonnable (typiquement moins de 10 multiplications-additions par échantillon de signal). Il faut cependant noter que préalablement à toute modification prosodique, il est nécessaire de déterminer la période du signal de parole (ainsi que son caractère voisé ou non voisé), opération qui est en général plus coûteuse que l'algorithme PSOLA lui-même ; en synthèse par concaténation d'unités, cela n'est pas très gênant, cette opération étant effectuée une fois pour toute lors de l'enregistrement des unités.

Des variantes de l'algorithme PSOLA, comme MBROLA ont connu également un grand succès pour le développement de synthèse multilingue. Une autre approche de modification non paramétrique est le vocodeur HNM (*Harmonic + Noise Model*) qui, par l'utilisation de la représentation sinusoïdale exploite la pério-



a le signal de parole original ainsi que les positions centrales des signaux à court-terme (en pointillés)

b les signaux à court-terme décalés (le pas de décalage inter-trame est multiplié par 1,25)

c le signal modifié obtenu par addition des signaux à court-terme décalés

Figure 5 – Modification de la fréquence fondamentale par un facteur 0,8 avec la méthode PSOLA

dicité de la parole dans le domaine fréquentiel plutôt que dans le domaine temporel comme PSOLA.

5.4 Synthèse par sélection et concaténation

Pour améliorer le naturel de la voix de synthèse, les systèmes de dernière génération utilisent des **unités de taille variable, plus grandes que les diphones** : l'idée est d'enregistrer un corpus de parole de grande taille et d'aller y puiser l'ensemble optimal de segments ou unités de synthèse. Ces unités peuvent être au choix, des segments de phrase, des mots ou des fragments de mots, des syllabes, des diphones, ou même des sons isolés. Dans un tel système, il existe ainsi de nombreuses possibilités pour le choix des segments pour une même chaîne phonétique. Ces approches sont ainsi couramment dénommées : synthèse par sélection dynamique d'unités acoustiques, ou synthèse par sélection/concaténation.

■ La **synthèse par sélection/concaténation est une extension de la synthèse par unités** concaténées, comme les diphones, ou la synthèse à parties manquantes. Le principe est simple, sans grande théorie, mais avec un souci d'efficacité : à partir d'une grande base de données de parole, contenant une ou plusieurs heures de signal, il s'agit de rechercher les plus longs segments contigus de diphones, de demi-phones, ou de segments plus petits, qui correspondent à la phrase à synthétiser. En ce sens, c'est une extension de la synthèse à parties manquantes, puisque l'on utilisera si possible des mots, voire des membres de phrase entiers. Mais comme le vocabulaire est illimité, il faut s'appuyer sur des unités comme les diphones, afin de compléter les énoncés si l'on ne parvient pas à trouver des tronçons de signal longs.

Étant donné une phrase, la première étape consiste comme pour les autres types de synthèse à transcrire son contenu phonétique, ainsi que des informations sur la constitution des énoncés : ponc-

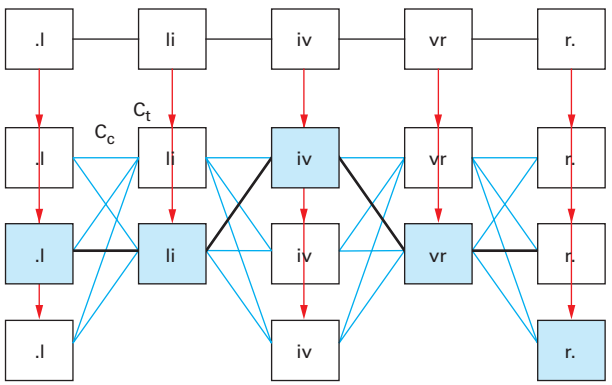
tuation, position des mots dans la phrase, des syllabes dans les mots, des syllabes accentuées par exemple. Ces informations vont permettre de rechercher à la fois des suites de phonèmes, mais aussi des segments qui partagent certaines propriétés prosodiques avec la phrase source.

■ La sélection de la suite optimale de segments dans la base est effectuée **à l'aide de fonctions de coût**. En général, deux fonctions de coût, le « coût de cible » et le « coût de concaténation » sont utilisées :

– **coût de cible** : il mesure et pondère l'avantage associé aux différents segments possibles, en termes de longueur maximale des segments, et de critères tels que la place du segment sélectionné dans la syllabe, le mot ou la phrase afin, d'obtenir une voix de synthèse avec une prosodie plus naturelle ;

– **coût de concaténation** : il mesure (en les pondérant) les quantités acoustiques (distorsion de concaténation, fréquence fondamentale moyenne) associées à la concaténation des unités.

Les unités optimales sont sélectionnées par une procédure d'optimisation des coûts, souvent avec un algorithme de programmation dynamique (ou un algorithme de Viterbi). La figure **6** illustre cette procédure sur un exemple simple pour la synthèse du mot livre.



(a) la chaîne de diphones à synthétiser, pour le mot isolé « livre »

(b) choix des unités dans la base

Tous les diphones correspondants sont recherchés, et le meilleur parcours dans le treillis de diphone est trouvé grâce à un algorithme d'optimisation, l'algorithme de Viterbi.

Figure 6 – Exemple de sélection des unités non uniformes
(algorithme type Viterubi)

Un exemple plus complet est donné dans le tableau **2**, le texte à synthétiser est « La chasse aux papillons », et l'algorithme de synthèse des segments de longueur variable puisés dans différentes phrases de la base de données de très grande taille.

■ Dans ce type de synthèse, la **prosodie n'est pas calculée explicitement**. Les segments sélectionnés sont longs, en général au moins des syllabes, parfois des mots ou des groupes de mots (les locutions courantes se retrouvent souvent entières : « bonjour », « c'est-à-dire » etc.). Ils sont porteurs de leur propre prosodie, qui est réutilisée directement, sans passer par un modèle.

L'algorithme de sélection va ainsi non seulement reconstituer la chaîne de phonèmes, mais aussi la structure prosodique (groupes de syllabes, début, milieu, fin de phrases, ponctuation, début, milieu, fin des mots, etc.). Cette prosodie implicite est en général

Tableau 2 – Exemple de sélection

« *La chasse aux papillons.* »

Phrases de la base de données	Segments de synthèse concaténés
/ .la Sas o papijO. /	
... d'autre part, **l**e ...	/ .l /
... **la cha**rte ...	/ la Sa /
... f**ace aux** voisins ...	/ as o /
... participe **au p**remier ...	/ o p /
... le **papy** boom	/ papi /
... publ**ié** ...	/ ij /
... ra**yon** chaud ...	/ jO /
... 1 000 chans**ons.**	/ O. /

Les segments de synthèse concaténés sont ceux extraits (texte en gras) des différentes phrases de la base de données.

d'un naturel étonnant, d'autant plus qu'elle offre beaucoup plus de variétés et moins de stéréotypes que la prosodie calculée par règles.

La **plupart des systèmes de synthèse commerciaux actuels sont basés sur la synthèse par concaténation**. La voix est souvent très naturelle et intelligible. Ce type de synthèse éprouve cependant des limites. C'est une synthèse assez coûteuse en termes de place mémoire, et la construction de bases de données de qualité demande beaucoup de soins et de temps. Le contrôle sur la voix synthétique est presque nul : c'est une méthode strictement non paramétrique, pour laquelle la qualité obtenue dépend uniquement du contenu de la base. Il n'est pas possible par exemple de changer la prosodie pour faire passer un contenu expressif, ou de changer la qualité de voix. Pour faire cela, il faut enregistrer et étiqueter une nouvelle base sonore. Cette synthèse est peu flexible : elle restitue bien ce qui est dans la base mais ne permet guère d'aller au-delà. Enfin, la qualité de synthèse peut être très variable d'une phrase à l'autre : dans certains cas parfaits, si les bons segments dans les bons contextes sont présents dans la base, dans d'autres cas très défavorables, si des combinaisons de segments manquent (ce qui est rare), ou si les contextes sont absents (ce qui est plus fréquent).

5.5 Synthèse paramétrique statistique

Une technique plus récente et en plein essor actuellement vise à **combiner la flexibilité de la synthèse paramétrique et la qualité de la synthèse utilisant des gros corpus de parole** : la synthèse paramétrique statistique.

Ce type de synthèse tire également avantage des dizaines d'années de recherche en reconnaissance de parole par des méthodes statistiques, et des outils disponibles dans ce cadre. La synthèse paramétrique statistique utilise le cadre général des **modèles de Markov cachés** (HMM pour *Hidden Markov Models*). Les travaux de Markov, au siècle dernier, ont montré que la succession des lettres dans les romans obéit à des lois de probabilité particulières. C'est l'origine des chaînes de Markov, ou automates probabilistes, particulièrement bien adaptées à la prédiction des enchaînements d'unités linguistiques, à tous les niveaux (phonèmes, syllabes, mots, suites de mots).

Contrairement aux approches par concaténation d'éléments sonores, la synthèse paramétrique par modèles statistiques repose

sur un modèle de signal et génerera un nouveau signal à travers un modèle de synthèse, pouvant être vu, en première approximation, comme le signal le plus probable, une moyenne d'un ensemble de signaux de parole similaires.

Tout comme la synthèse par règles, la synthèse paramétrique statistique repose sur un **modèle paramétrique du signal**, un modèle source-filtre. Mais, contrairement à la synthèse par règles, les paramètres de ce modèle ne sont pas analysés et définis explicitement par un expert, et de façon déterministe. Ils sont appris grâce à des modèles statistiques, sur un gros corpus de parole, sous forme de chaînes de Markov, et générés de façon probabiliste (la succession la plus probable de paramètres est émise) étant donné un texte d'entrée à synthétiser. Ce principe est illustré sur la figure **7**.

■ La **construction d'un système de synthèse** consiste, tout comme en synthèse par sélection-concaténation, à enregistrer un gros corpus de parole, et à l'étiqueter. Les étiquettes utilisées sont du même type composite que pour la synthèse par concaténation : phonème, phonèmes adjacents à droite et à gauche, syllabe, position du phonème dans la syllabe, de la syllabe dans le mot, mots adjacents, ponctuation, catégories syntaxiques, etc. Tout type d'étiquette que l'on peut apposer sur des segments de façon régulière est susceptible d'un apprentissage statistique, par comptage des occurrences.

Le signal de parole enregistré est analysé par un modèle source-filtre, avec d'un côté les paramètres du filtre, sous forme en général non pas de formants mais de paramètres spectraux plus élaborés, et de l'autre les paramètres de la source, c'est-à-dire les paramètres prosodiques (intonation, durée, qualité vocale, intensité).

Les paramètres spectraux utilisés pour représenter le filtre sont ceux utilisés en reconnaissance de parole. En effet, les synthétiseurs par HMM sont directement issus des outils robustes et sophistiqués développés pour la reconnaissance depuis plusieurs décennies (par exemple, les MFCC, *Mel Frequency Cepstral Coefficients*, leurs dérivées et dérivées secondes). Ces paramètres se prêtent bien à l'apprentissage automatique et leur extraction est très robuste, contrairement aux formants par exemple.

Ainsi, lors de la phase d'apprentissage, des modèles probabilistes représentant la base de données sonore sont construits. La base de données elle-même n'est pas conservée dans le système de synthèse, mais seulement sa représentation, ainsi l'empreinte mémoire du système est réduite.

■ Au **moment de la synthèse**, comme pour les autres modes de synthèse, la phrase d'entrée est analysée par le module linguistique et enrichie d'étiquettes linguistiques nombreuses. Ces étiquettes sont transmises aux modèles de Markov cachés (sous-jacents) construits lors de l'apprentissage, qui génèrent les paramètres les plus probables.

Par exemple, les modèles de phonème génèrent les paramètres spectraux les plus probables du phonème en fonction de son identité et de son contexte (phonèmes voisins, syllabe, mot, etc.). Le modèle d'intonation génère des points cibles pour les phonèmes voisés en fonction des étiquettes d'entrée. Les durées sont générées par d'autres modèles statistiques appris sur le corpus.

Lorsque les paramètres ont été générés, ils pilotent un vocodeur de synthèse qui génère le signal acoustique proprement dit. Plusieurs types de vocodeurs différents existent dans les différents systèmes.

Cette approche est très prometteuse et commence à rivaliser en termes de qualité avec les systèmes par sélection/concaténation même si elle ne donne pas encore lieu, à ce jour, à des systèmes commerciaux d'une qualité comparable à ceux par concaténation. En effet, l'utilisation d'un vocodeur entraîne toujours une dégradation de qualité des échantillons de parole. Par contre, la flexibilité totale et les progrès des modèles statistiques, grâce à la reconnaissance de parole rendent ces systèmes très attractifs.

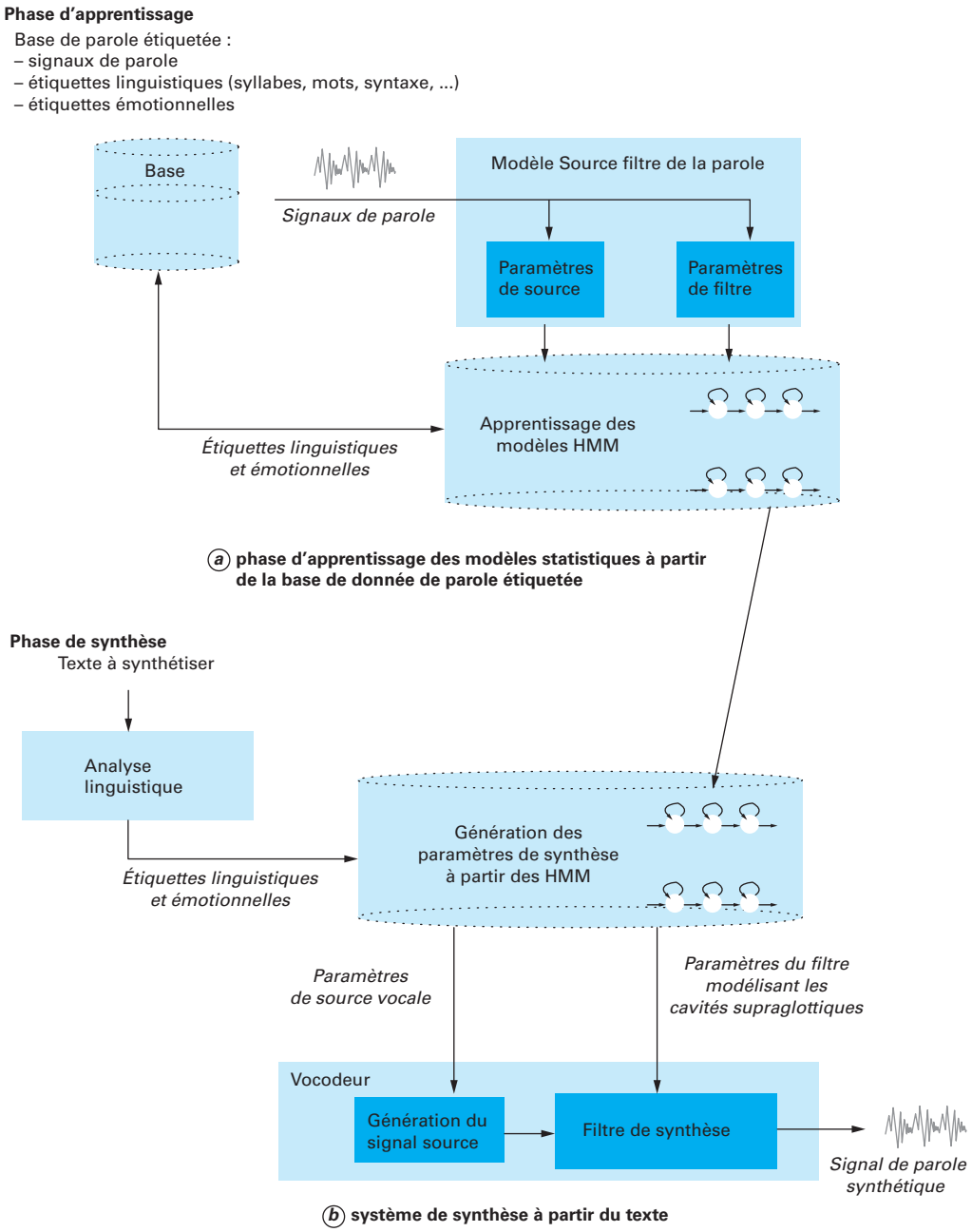


Figure 7 – Schéma de principe de la synthèse paramétrique statistique

5.6 Construction du corpus textuel et sonore

Pour la synthèse par sélection/concaténation, comme pour la synthèse paramétrique statistique, **la taille de la base de données est nécessairement importante**. Au moins une heure de parole (à comparer avec les 2 ou 3 min d'une base de diphones), parfois beaucoup plus, semble incontournable pour couvrir l'ensemble des combinaisons possibles de phonèmes et de contextes prosodiques.

La construction de la base de données est un sujet de recherche en soi. En effet, il faut qu'elle soit d'une taille optimale, c'est-à-dire que la plupart des segments qu'elle contient soient potentiellement utilisés, mais aussi qu'elle soit aussi complète que possible.

La base doit bien entendu contenir tous les phonèmes, tous les diphones, mais aussi des contextes plus longs, si possible toutes les syllabes. En plus de ces éléments phonétiques, l'algorithme de sélection utilise dans le calcul des coûts des éléments syntaxiques ou prosodiques, comme la place de la syllabe dans le mot, la place du mot dans la phrase, la ponctuation, la catégorie du mot, la longueur de la phrase, la durée de la syllabe, la valeur de fréquence fondamentale, etc.

Ainsi, la segmentation et l'étiquetage de cette base est réalisée à l'aide d'outils de reconnaissance vocale qui permettent d'obtenir des marques de frontières entre les différentes unités de base (certains systèmes utilisent des unités de base très petites comme le demi-phone). Des outils d'analyse linguistique et d'analyse du signal complètent la construction de la base. Cette automatisation permet de construire des nouvelles voix de synthèse assez simplement puisqu'il suffit pour cela d'enregistrer un nouveau locuteur (ou nouvelle locutrice).

■ La **qualité de synthèse** obtenue dépend principalement de celle du corpus sonore enregistré. La construction de ce corpus est donc fondamentale. Ces méthodes permettent d'ailleurs de reconstituer en synthèse de parole la voix d'un locuteur pour lequel on possède beaucoup d'enregistrements, comme celle d'un homme politique, d'un acteur, voire sa propre voix en s'enregistrant soi-même. La qualité de la synthèse dépend également du locuteur : même si exactement les mêmes outils et les mêmes phrases pour la base de données ont été utilisés, deux locuteurs différents ne donneront pas des systèmes de même qualité. Cela dépend de la netteté d'articulation, de la constance de prononciation, de la résistance à la fatigue vocale pour de longues séances d'enregistrement, et de la sonorité même de la voix.

6. Applications de la synthèse de parole

6.1 Exemples d'applications

De nombreuses applications commerciales intègrent des systèmes de synthèse de parole. À l'heure actuelle, le marché principal de ce type de technique est celui des services de télécommunications. Ces services constituent l'exemple typique de situations dans lesquelles la synthèse de parole est le seul moyen par lequel un système informatique peut transmettre des informations à ses utilisateurs. Parmi les applications de la synthèse à partir du texte dans le domaine des services de télécommunications, citons :

– les services de réservation ou de prise de commandes téléphoniques ;

– les services d'information téléphonique pour lesquels le recours à la synthèse de parole se justifie, surtout lorsque l'information est amenée à évoluer vite, ce qui est notamment le cas pour les services bancaires (avec la fourniture, entre autres de l'état des comptes), les annonces météorologiques et routières, la lecture de méls ou de pages Internet. La synthèse de parole est aussi utilisée dans des contextes où le nombre des réponses potentielles du système est très important comme dans les applications de renseignements téléphoniques ;

– les majordomes, assistants personnels, pour les téléphones mobiles ou autres terminaux, qui peuvent lire des messages reçus ou des courriers électroniques ;

– une application ambitieuse est envisagée à l'heure actuelle avec la téléphonie interprétée qui devrait permettre à deux corres-

pondants ne parlant pas la même langue de dialoguer par téléphone. Cette application fait intervenir plusieurs des grandes problématiques du traitement de la parole – reconnaissance, synthèse –, et bien sûr traduction automatique.

La synthèse de parole est aussi couramment employée dans des situations où l'utilisateur d'un système informatique n'a pas le loisir de consulter un écran, ou bien en complément de l'écran (cabine de pilotage d'un avion, systèmes industriels de fabrication, appareillage médical, etc.). Dans ce type d'applications, le rôle de la synthèse de parole consiste principalement à faire passer des informations brèves comme les messages d'erreurs du système. Les applications dans les systèmes d'information, fixes ou mobiles sont également nombreuses :

– portail vocaux d'application libre service ou de sites Internet ;
– systèmes de navigations ;
– systèmes de renseignement ;
– accessibilité des services ;
– vocalisation de journaux et de livres électroniques ;
– lecteurs d'écran ;
– jouets, robots et autres systèmes embarqués ;
– jeux vidéo ;
– jeux sérieux, éducation, *edutainment*.

La qualité accrue des systèmes de synthèse permet maintenant de développer des applications d'apprentissage des langues étrangères qui deviendront les évolutions naturelles des applications actuelles de dictionnaire électronique de poche avec leur capacité à synthétiser des mots ou des phrases dans plusieurs langues.

Un autre aspect important des applications de la synthèse de parole à partir du texte concerne les services pour personnes handicapées. Dans ce domaine, le couplage de la synthèse de parole avec les techniques de reconnaissance automatique de caractères a permis la mise au point de véritables « machines à lire » pour les mal ou non-voyants.

6.2 Interfaces de programmation

La synthèse de parole est intégrée dans les systèmes d'exploitations, les bibliothèques de logiciels ou les services sous la forme d'interface de programmation (API : *Application Programmation Interface*), comme par exemple le *Speech API* (SAPI) de Microsoft.

Le *World Wide Web Consortium* (W3C) recommande le protocole SSML (*Speech Synthesis Markup Language*), protocole basé sur XML, afin d'annoter ou d'enrichir la synthèse avec des marqueurs prosodiques comme la fréquence fondamentale, le débit, les pauses, le volume, etc.

6.3 Produits

L'offre en produits commerciaux de synthèse est aujourd'hui répandue. La plupart de ces systèmes sont multilingues, c'est-à-dire capables de produire des voix de synthèse dans plusieurs langues différentes. Ces systèmes incluent tous la synthèse de l'anglais (généralement, américain), et de la plupart des langues européennes et des grandes langues orientales.

Les configurations logicielles diffèrent suivant le type de produits et les applications. La plupart du temps cependant, l'obtention d'une voix de synthèse ne nécessite plus un matériel spécifique (si ce n'est une carte de restitution du son, disponible en standard sur à peu près toutes les plates-formes), la synthèse proprement dite ne requérant en fait qu'une fraction de la puissance de calcul d'un processeur moderne. Pour certaines applications spécifiques (serveurs vocaux ou applications embarquées), des implantations matérielles sont encore souvent nécessaires.

Comme pour beaucoup de produits dans le domaine des technologies de l'information et de la communication, les offres

évoluent très vite, les compagnies fusionnent ou sont rachetées, et le paysage est plutôt instable, à part pour les structures de grande taille.

Le paysage actuel est caractérisé par la concentration de l'offre commerciale sur peu d'acteurs. On note aussi la disparition des compagnies de téléphonie dans l'offre de produit, au profit de petites *startups* qui en sont issues, et des compagnies de logiciel. Nous donnons ci-dessous une liste non exhaustive des acteurs du domaine et fournissons pour certains d'entre eux une description plus détaillée de leur offre.

Nota : se reporter à la partie documentation de cet article pour une présentation plus détaillée des acteurs commerciaux des produits de synthèse.

7. Évaluation de la synthèse

7.1 Boîte noire ou boîte de verre

L'évaluation de la synthèse est nécessaire pour comparer les systèmes entre eux et pour mesurer les progrès réalisés. L'évaluation de la parole de synthèse peut être globale (le signal généré) ou porter sur un des aspects du système.

L'évaluation globale est celle en général du client, qui écoute et évalue le synthétiseur sur la parole produite. L'évaluation des composantes intéresse plus le chercheur ou le développeur, qui cherche à améliorer un aspect du système.

La synthèse à partir du texte est en effet une chaîne de traitements, depuis le texte jusqu'au signal acoustique. C'est donc le maillon le plus faible de la chaîne qui en limitera la qualité, d'où l'importance d'évaluer chaque module (évaluation analytique ou interne, ou « boîte de verre »), en plus de l'évaluation globale (évaluation externe, ou « boîte noire »). En général, la référence de l'évaluation des systèmes est la parole naturelle, ou de la parole naturelle avec un certain niveau de bruit.

L'évaluation de la parole peut porter sur les aspects suivants :

1. Évaluation globale :

a. intelligibilité (netteté à l'aide de syllabes, phrases dépourvues de sens, parole téléphonique, etc.),

b. qualité globale, suivant plusieurs critères (test d'opinion moyenne : *Mean Opinion Score* ou MOS) ;

2. Évaluation analytique :

a. transcription graphème-phonème (taux de transcription correct, en particulier pour les noms propres),

b. prosodie (expressivité, agrément, etc.),

c. vocodeur – synthétiseur acoustique (qualité du signal, artefacts).

7.2 Évaluation de qualité globale

Pour la qualité globale, une procédure multidimensionnelle d'évaluation a été normalisée en 1994 par l'Union internationale des télécommunications (UIT-T P.88). Les échantillons de parole doivent durer de 10 à 30 secondes, et il est recommandé d'utiliser une référence de parole naturelle (éventuellement dégradée).

Les sujets écoutent chaque échantillon deux fois, pour une durée d'environ une heure (par exemple quatre stimuli pour quatre systèmes et trois références), en incluant les instructions et l'apprentissage.

Les huit dimensions d'analyse sont :

1. Acceptabilité (pensez-vous que cette voix convienne pour tel service ?) ; 1 : oui, 2 : non.

2. Impression d'ensemble (comment évaluez-vous la qualité de ce que vous venez d'entendre ?) ; 1 : excellente, 2 : bonne, 3 : correcte, 4 : faible, 5 : mauvaise.

3. Effort d'écoute (comment décririez-vous l'effort nécessaire pour comprendre le message ?) ; 1 : relaxation complète, aucun effort nécessaire, 2 : attention nécessaire, pas d'effort notable, 3 : effort modéré nécessaire, 4 : effort nécessaire, 5 : rien de compréhensible, quel que soit l'effort.

4. Compréhension (avez-vous trouvé des mots difficiles à comprendre ?) ; 1 : jamais, 2 : rarement, 3 : parfois, 4 : souvent, 5 : constamment.

5. Articulation (peut-on distinguer les sons avec netteté ?) ; 1 : oui, très net, 2 : oui, assez net, 3 : relativement net, 4 : non, pas très net, 5 : non, pas du tout net.

6. Prononciation (avez-vous remarqué des anomalies de prononciation ?) ; 1 : non, 2 : oui, mais pas gênantes, 3 : oui, gênantes, 4 : oui, très gênantes.

7. Débit de parole (que pensez-vous de la vitesse moyenne de prononciation ?) ; 1 : beaucoup trop rapide, 2 : trop rapide, 3 : correcte, 4 : trop lente, 5 : beaucoup trop lente.

8. Agrément de la voix (comment décririez-vous cette voix ?) ; 1 : très agréable, 2 : agréable, 3 : correcte, 4 : déplaisante, 5 : très déplaisante.

Le test d'opinion moyenne MOS est le plus souvent utilisé pour évaluer la synthèse. Ce test utilise généralement une échelle de cinq points, entre « très mauvaise qualité » et « excellente qualité ». La parole naturelle obtient généralement des scores de l'ordre de 4,5, et la parole de synthèse actuellement des scores de l'ordre de 3,5 pour les meilleurs systèmes, dès lors que les phrases sont longues.

8. Conclusion

8.1 Bilan

Bien que la liste des applications actuelles des systèmes de synthèse de parole soit assez conséquente, il serait faux de croire que celle-ci constitue une technique entièrement maîtrisée. Cependant, les travaux de recherche menés depuis des années ont permis d'atteindre une qualité qui se rapproche de celle de la voix naturelle.

La **qualité de synthèse est un problème crucial**. Il se manifeste essentiellement par le fait que la compréhension de la parole synthétique exige, de la part de l'auditeur, un effort plus important que pour la parole naturelle. Cet effort supplémentaire est rendu nécessaire par les artefacts éventuels du traitement, les erreurs de prononciation, la prosodie insuffisamment expressive, ou de manière plus générale, par le manque de naturel de l'élocution. Actuellement, si la qualité des systèmes de synthèse par sélection d'unités non uniformes est suffisante pour de nombreuses applications (lecture de méls, de messages d'information météo ou de navigation...), elle reste encore trop faible pour permettre une utilisation « prolongée » de la parole de synthèse ou pour des tâches véritablement expressive (lecture de livres par exemple).

Un exemple du type de problèmes qui restent associés à la synthèse de parole est le pourcentage important de prononciation incorrecte pour les noms propres. Ce problème n'est pas totalement inconnu dans le cas de la parole naturelle (il est même familier pour les enseignants confrontés à l'épreuve de l'appel en début d'année) ; toutefois, un locuteur humain est capable d'éliminer une proportion importante des erreurs potentielles en faisant appel à ses connaissances culturelles (notamment à celles qui concernent l'origine géographique du nom).

8.2 Perspectives

Les travaux en cours s'attachent ainsi à améliorer la flexibilité des systèmes suivant plusieurs axes :

– variabilité de la voix de synthèse au cours du temps, en fonction de l'énoncé ;

– possibilités d'expressivité accrue (attitudes, émotions, personnages particuliers) ;

– création rapide de nouvelles voix de synthèse.

La décennie passée a vu le développement de nouvelles techniques de synthèse (par **méthodes statistiques paramétriques**) proches de celles utilisées en reconnaissance de la parole. Ces techniques ne sont pas encore passées dans les produits mais sont très près de franchir ce pas.

Les offres de produit se sont diversifiées, avec en particulier le déploiement d'applications sur toutes les plates-formes, du serveur distribué au mobile, la possibilité de « studios » de synthèse pour préparer des messages d'une qualité équivalente à celle d'un locuteur enregistré, le développement d'assez nombreuses langues.

Le principal enjeu est aujourd'hui de rendre la synthèse interactive, capable de répondre et d'évoluer avec son environnement, capable d'exprimer des affects sociaux ou des émotions, et d'employer toute la gamme expressive de la voix humaine. Pour cela, des progrès sur la modélisation acoustique aussi bien que sur la compréhension des textes seront nécessaires.