

Mémoire de fin d'études : Traduction
instantanée et émotions

Florian Constant et Aubry Gudel

24 février 2014

Table des matières

1	Introduction à la problématique	5
1.1	Le scénario	5
1.2	Les composantes	6
2	Résolution du système	9
2.1	Reconnaissance automatique de la parole	9
2.1.1	Fonctionnement générale	9
	Étude du signal (décodage acoustico-phonétique) . . .	10
	Reconstruction sémantique	12
2.1.2	Et pour nous ?	12
2.2	Traduction	13
2.2.1	Présentation et but	13
2.2.2	Solutions	14
	Traduction statistique	16
	Traduction par règles	18
	Traduction par l'exemple	19
2.2.3	Notre choix	21
2.3	Synthèse vocale	21
2.3.1	Présentation et but	21
2.3.2	Principes générale	21
	Traitement automatique du langage naturel	22
	Synthétiseur vocale	23
2.3.3	Dans notre cas	25
2.4	Émotions	26

2.4.1	Présentation et but	26
2.4.2	Identification des émotions	27
	Quels sont les facteurs d'identification des émotions?	27
	Comment allons-nous les identifier?	30
2.4.3	Retranscription des émotions	31
	Retranscription par synthèse vocale	32
	Limites du système	34
3	Aller plus loin	35
3.0.4	Au niveau de la détection des émotions	35
3.0.5	Émotions et traduction	36
3.0.6	Synthèse de l'émotion	36

Résumé

Nous proposons ici une étude sur la traduction instantané voix vers voix dans le cadre d'une réunion. L'objectif principale sera de proposer une solution viable dans l'environnement d'une salle de réunion et à l'aide des technologies disponible. Nous aborderons aussi plus en détail la gestion des émotions dans le processus : nous pensons en effet que leur interprétation et restitution est un élément clef dans l'élaboration d'un système voix vers voix pour permettre une plus fine compréhension du locuteur dans le court laps de temps qui est imposé par la dynamique d'une conversation.

Nous traiterons donc dans un premier temps de l'ensemble des modules nécessaires à la traduction voix vers voix avant de nous attarder sur la prise en compte des émotions et les limites actuels.

Abstract

In this work we aim at exploring speech to speech translation in real time in the closed environment of a meeting room. The main goal is to propose a viable solution fitting a meeting room's conditions with today's technologies. Next we will adress how emotions can be handled by the system and how processing emotions can give an edge when trying to understanding what is being said in such a fast paced situation as a conversation.

To that end we will adress the basic modules needed for a speech to speech translation in real time, then we will study how emotions can be incorporated in the system and what are todays limitations in their handling.

Mots clés

Traduction, Synthèse vocale, Émotion, Traduction instantanée, Retranscription, SMT, RBMT

Translation, Speech, Emotion, Real-time translation, transcript, SMT, RBMT

Chapitre 1

Introduction à la problématique

1.1 Le scénario

Avant toute chose, nous allons commencer par définir le scénario, sans quoi, il serait difficile de se poser des limites dans nos réflexions.

Notre scénario

Dans un cadre professionnel, dix personnes de nationalités différentes et donc ne parlant pas la même langue se joignent autour d'une table afin d'avoir une réunion. Chaque personne parlera dans sa langue maternelle.

Afin que tout le monde puissent comprendre ce que les autres disent et que la réunion soit dynamique, un système sera mis en place afin de traduire chaque personne et de retranscrire vocalement la traduction. Ce système sera relativement poussé puisqu'il traduira en temps réel, c'est-à-dire qu'il fera la traduction alors même que l'interlocuteur n'a pas fini sa phrase. De plus, ce système sera capable d'analyser les émotions dans la voix et de la restituer dans la synthèse vocale.

Nous allons devoir identifier chaque composante de ce système dans notre scénario et affiner celui-ci.

1.2 Les composantes

Le contexte

Premièrement, nous allons devoir déterminer le contexte de la future discussion pour notre système. En effet, que le système sache dans quelle contexte se situe la discussion est très important car il permettra d'améliorer grandement la fiabilité de la traduction. Lors de la traduction le niveau de langage approprié pourra être choisi, ce qui rendra la traduction plus fidèle.

On pourrait imaginer que le système analyse la discussion afin de déterminer ce contexte mais le système serait moins efficace le temps de cette analyse. C'est pourquoi nous allons partir du principe que quelqu'un initialise le contexte avant la réunion (via une interface par exemple). Le contexte choisi par cette personne sera donc « Professionnel ».

Identification et langues

Une fois le contexte choisi, la réunion peut commencer. Les personnes commenceront alors à parler. Ici nous trouvons alors un autre problème : Comment détecter qui parle ? Et dans quelle langue ?

Afin de répondre à la première question, nous avons pensé à un système qui analyse les fréquences des voix ainsi que leurs signatures afin d'identifier chaque personne si le système avait un micro central. Cela fonctionnerait certainement mais dans ce cas, comment restituer dans la langue de la personne la traduction sans avoir une cacophonie de traduction dans la salle ? Il nous paraît alors évident qu'il soit nécessaire que chaque participant ait une oreillette dans laquelle on lui traduit les discussions.

Afin d'éviter un processus d'identification des propriétaires par rapport à leur voix afin de ne pas leur traduire leur propre parole, il nous a paru plus simple de supposer que les oreillettes contiennent un micro dans lequel parle

l'interlocuteur. De cette manière, chaque personne sera identifiée de manière simple et sûre.

Pour ce qui est de la seconde question, c'est-à-dire de comment connaître la langue, il s'agira de la détecter en fonction de ce que la personne dira. Un peu à la manière d'un Google Translate avec l'écrit. Mais nous reviendrons sur la réalisation de cette tâche plus tard.

Reconnaissance automatique de la parole

À ce stade, il nous faut encore réussir à mettre sous format textuel ce qu'une personne dit afin de le traduire par la suite. Il existe des logiciels spécialisés dans ce travail, on les appelle des logiciels de Speech To Text, ou en français : « Reconnaissance automatique de la parole ». Il existe une multitude de logiciels comme ceci, c'est pourquoi nous n'allons pas plus en parler pour le moment et estimons que ce travail est fait.

Traduction

Maintenant que l'on dispose du texte, nous pouvons enfin le traduire dans toutes les autres langues présentes dans la salle. Pour ce faire nous allons devoir utiliser un traducteur, seulement, pour parfaire la traduction, le traducteur devra prendre en compte le contexte de la discussion afin de choisir les traductions les plus adaptées.

Synthèse vocale

Il est désormais temps de restituer la traduction à chaque interlocuteur. Pour réaliser cela, nous allons devoir faire appel à un système de synthèse vocale. Une synthèse vocale a pour but d'exprimer un texte par la parole, de la manière la plus naturelle et humaine possible dans notre cas. De nos jours, il existe une multitude de logiciels de synthèse vocale, nous en parlerons plus en détails dans une autre partie.

Émotions

Il temps de parler d'une composante qui est transverse à plusieurs autres, citées plus haut. Il s'agit de l'émotion, nous souhaitons que la traduction permette à celui qui écoute de ressentir l'émotion transmise par la personne qui parle. Pour se faire, il va falloir analyser la voix (Speech To Text), en analyser les émotions puis la retranscrire dans le choix de certains mots peut-être, durant la traduction. Puis il faudra la retranscrire via la synthèse vocale de la manière la plus fidèle possible.

La composantes de l'émotion est très intéressante car encore absente des solutions temps réel actuel, sa détection et synthèse en temps réel est un problème toujours non résolu alors que son apport en terme de dynamisme comme d'aide à la compréhension -qui sont deux facteurs très important pour un système temps réel- sont très importants. C'est pourquoi nous nous pencherons plus particulièrement sur ce facteur et détaillerons comment la détecter, comment la comprendre et comment la retransmettre dans les limites imposés de notre scénario.

Dans un premier temps nous aborderons les solutions existantes permettant de répondre à chacune des étapes de notre problématique avant de faire un point plus détailler sur la reconnaissance et la retranscription des émotions. Pour finir nous évoquerons les limitations inhérente à notre solution.

Chapitre 2

Résolution du système

2.1 Reconnaissance automatique de la parole

Introduction

La reconnaissance automatique de la parole (que nous abrègerons ci-après en RAP) est un système qui permet d'analyser la parole d'une personne pour la retranscrire sous la forme d'un texte.

Cette technique est fréquemment utilisé de nos jours. Que ce soit dans les serveurs vocaux (messageries, assistance...), sur un ordinateur avec la dictée vocale ou dans les assistants personnels (Google Now, Siri).

2.1.1 Fonctionnement générale

Petit historique

Les premiers systèmes de RAP numériques datent d'il y'a plus de cinquante ans, ces premières solutions étaient limitées à la reconnaissance de mots isolés dans un vocabulaire très limité (cardinal d'une dizaine de mots). Nous nous intéresserons ici uniquement aux techniques modernes permettant de couvrir une langue complète et de reconnaître de la parole continue, soit des phrases et discours complets ayant un sens.

Parole continue

Les système de reconnaissance de parole continue reçoivent en entrée un signal (analogique ou numérique) correspondant à un message orale à re-transcrire en texte, ce qui pose un certain nombre de difficultés successive. La continuité de texte impose au système de :

- redécouper le signal en mots, tâche compliqué par le fait qu'à l'oral il n'y a pas de réel séparation (pause) entre les mots, si ce n'est pour ponctuer, et que dans certain cas des liaisons sont faite entre plusieurs mots une syllabe pouvant ainsi se retrouver *à cheval* entre deux mots
- rassembler les mots en phrase en suivant des problématiques de syntaxe
- de vérifier que le résultat obtenue est sémantiquement correcte, chaque phrases doit avoir un sens

Le modèle le plus courant actuellement consiste à séparer les deux types principaux de difficultés, d'une part le signal acoustique est décodé vers une information phonétique puis les informations phonétiques sont traité par des algorithmes de modélisation du langage.

Tout d'abord le décodage phonétique s'opère en analysant par fenêtre le signal d'origine puis en comparant ses composés à la bibliothèque d'unités connue utilisée par le système (mot, syllabe, diphone, phonème, etc.). Dans un second temps les informations décodé doivent être traité pour obtenir des phrases syntaxiquement et sémantiquement correcte, pour ce faire les systèmes actuels reposent sur des procédés statistiques étudiant la probabilité d'une suite de mots et ajustant ainsi les séquences trouvé.

Étude du signal (décodage acoustico-phonétique)

L'étude du signal s'opère le plus le plus souvent via une analyse paramétrique de ce dernier et permet ainsi de retrouver l'élément acoustique le plus probable correspondant au signale.

Lors de l'acquisition des données acoustiques des pertes liées au matériel (plage de fréquence du microphone) sont inévitable ainsi que des perturbations liées à l'environnement (réverbération de la salle, bruit ambiant). Nous n'étudierons pas ici les différentes solutions qui existent permettant d'amoindrir l'impacte des deux facteurs précédemment énoncé et passerons directement à l'étude du signal nettoyé.

L'étude se fait alors par fenêtrage, le signal de la parole n'évoluant que peu sur des durées de quelques millisecondes (stationnarité locale). Sur chaque fenêtre une analyse spectrale est alors effectuée, par exemple grâce à la décomposition du signal sur la fenêtre considéré via une transformée de fourier puis en analysant le résultat avec un ensemble de filtres passe-bande permettant la représentation de l'échantillon par un sonagramme. Il Suffirait alors d'effectuer une reconnaissance de forme en comparant le résultat à notre bibliothèque et en conservant l'élément le plus probable.

Malheureusement il n'est pas possible d'effectuer directement une telle opération, la durée de chaque unité à reconnaître n'étant pas stable il faut pouvoir faire coïncider le signal relevé avec les données de notre bibliothèque via des procédés de normalisation temporelle.

Pour ce faire il existe plusieurs modèles notamment les modèles Markoviens Cachés, les modèles neuronimétriques ou encore les algorithmes de comparaison dynamique.

La technique la plus couramment utilisée aujourd'hui se fait par modèles de Markov caché. Il est alors possible de retrouver l'ensemble des états probablement parcouru à partir des observations faites en isolant le chemin le plus probable.

Reconstruction sémantique

Une fois les unités phonétiques de bases isolées il reste alors à identifier les formes lexicales correspondantes. Cette étape est appelé le décodage. Les techniques de décodages reposent sur des graphes on retrouve alors des résolutions se basant sur l'algorithme A^* . On génère alors un graphe où les nuds comportent tous les mots probables composés par les unités phonétique préalablement isolés.

2.1.2 Et pour nous ?

Dans le cadre de notre scénario, plusieurs des difficultés peuvent être purement ignorées. Ainsi l'étape de soustraction du bruit ambiant n'est pas nécessaire, les réunions se déroulant typiquement dans des lieux éloignés des perturbations sonores.

Pour ce qui est du problème de parole concurrente (deux interlocuteurs se mettent à parler en même temps), celui ci pose deux problèmes : séparer ce que disent les deux intervenants pour permettre d'identifier ce que chacun dit, puis retranscrire le résultat à l'auditoire. Si il est certainement possible de séparer ce que dit chaque intervenant cela demanderait des traitement supplémentaire qui ne sont d'après nous pas justifiable : comment retranscrire ensuite à l'auditoire ce qui est dit ? Le résultat de chaque traduction ne peut pas être rendu en même temps, aucun ne serait alors intelligible et si l'on décale la synthèse de l'une des traduction quand la passer ? Elle se retrouverait déplacé à plus tard indéfiniment à mesure que d'autres intervenants prendraient la parole. Nous avons donc trouvé préférable d'ignorer ce cas en nous reposant sur la discipline des participants.

Quand aux techniques de reconnaissances se posent trois questions : l'unité à utiliser, le modèle de décodage acoustico-phonétique et l'algorithme de reconnaissance lexical/grammatical ; ce dernier étant directement liée au choix du modèle de décodage acoustico-phonétique.

Dans le cas de la parole continue il a été identifié que l'unité la plus adaptée était le phonème : en effet l'ordre de grandeur des phonèmes nécessaire à la modélisation d'un langage est très faible, même lors de l'utilisation de phonème contextualisé (prenant en compte de possibles liaisons) et permet ainsi une relative légèreté du dictionnaire comportant les mots reconnaissable. Nous utiliserons donc le phonème.

Les autres paramètres n'ayant pas d'impact particulière sur le cur de notre recherche nous sélectionnerons le système le plus utilisé, les Modèles de Markov cachés, celui-ci ayant déjà prouvé son efficacité.

2.2 Traduction

2.2.1 Présentation et but

La traduction est le principe de faire passer un texte d'une langue à une autre. Elle sert à représenter un texte dans une autre langue, elle en garde donc le même sens et les deux textes comportes donc beaucoup de similitudes.

Historiquement, la traduction à d'abord été un travail d'humain. Puis par la suite, c'est l'informatique qui s'en est chargé. Lorsqu'un système informatique traduit un texte, nous appelons cela de la « traduction automatique ». Dans notre scénario, c'est cette traduction automatique dont nous allons parler.

La traduction ne prend pas seulement en compte le fait de traduire une suite de mot d'une langue à une autre. Il faut préserver le sens et être le plus fidèle possible. Pour cela, le traducteur (ou système de traduction) doit connaître le contexte du texte, la grammaire des deux langues, ainsi que leur cultures. En effet, la culture compte pour beaucoup car certains mots dans une langue n'ont pas la même signification dans une autre, il faut alors transformer les mots afin de refléter l'idée du texte original. C'est pourquoi

les traduction les plus fiables et fidèles sont faites par des humains spécialisés dans la traduction.

2.2.2 Solutions

Tout d’abord, il faut savoir que dans la traduction non informatisée il y a trois phases durant la traduction :

- La compréhension : comprendre le sens du texte d’origine
- La dé-verbalisation : garder le sens du texte sans les mots
- La ré-expression : formulation du sens du texte dans la langue d’arrivée

Dans la traduction informatique, la compréhension sera appelée **analyse**, la dé-verbalisation **transfert** et la ré-expression **génération**. Ces trois phases sont représentées dans le triangle de Vauquois ci-dessous.

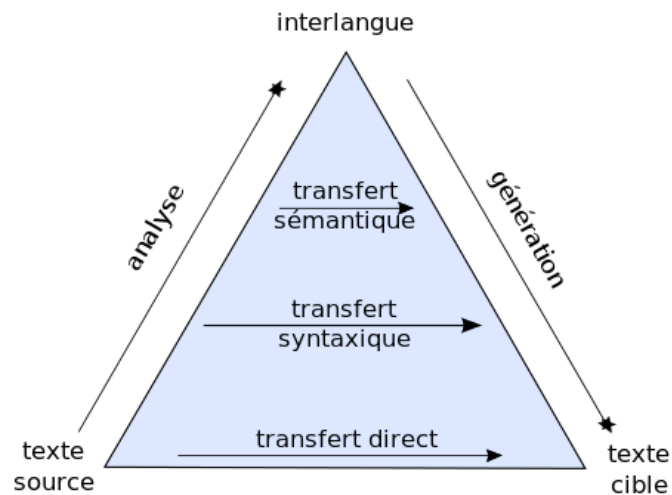


FIGURE 2.1 – Triangle de Vauquois

Ce schéma est représentatif des différents chemins possibles afin de traduire un texte source dans une langue cible. Ces différents chemins visibles dans le schéma sont les différentes manières de faire actuellement utilisées. Nous

pouvons interpréter cela comme « Plus l'analyse est longue, plus le transfert est court ».

Il existe quatre possibilités :

- **Le transfert direct** : il n'y a pas presque pas d'analyse, toute la traduction se joue au niveau du transfert. Les méthodes de traductions utilisant ce type de transfert sont la traduction par l'exemple et la traduction statistique. Dans cette possibilité, la traduction n'est qu'un processus de décodage.
- **Le transfert syntaxique** : ici, le transfert est syntaxique, c'est à dire que l'on va s'appuyer sur les arbres syntaxique des langues afin de construire des phrases de même sens. La méthode de traductions utilisant ce type de transfert est la traduction automatique à base de règles.
- **Le transfert sémantique** : le transfert est sémantique, c'est à dire que l'on traduit à partir du sens du texte d'origine. Il s'agit de la méthode de traduction que nous utilisons (les humains). Il n'y a que très peu de traducteur automatique se basant sur le transfert sémantique, de part sa complexité à modéliser les sémantiques ainsi que part sa difficulté à la mettre en place.
- **L'interlangue** : cette possibilité supprime le transfert. La traduction devient alors universelle. Et il ne reste plus que les phases d'analyse et de génération. On désigne également linter-langue sous le nom de « langue pivot ». Cette méthode n'a pas eu de succès chez ceux qui l'ont essayé.

De nos jour, les moteurs de traduction utilisent pour la plupart la traduction par règles ou la traduction statistiques. Mais il existe une approche utilisant ces deux systèmes de traductions qu'utilise les leaders du marché que sont Systran, Google Translate et Bing Translator.

Traduction statistique

La traduction automatique statistique est basée sur l'utilisation de « modèles statistiques » auto construits à partir de corpus monolingues et bilingues. La construction des « modèles statistiques » est rapide mais requiert d'avoir à disposition des volumes importants de textes traduits. Généralement, un modèle bilingue nécessite au minimum 2 millions de mots pour la traduction dans un domaine spécifique, mais il en nécessite beaucoup plus pour le domaine général. La traduction automatique statistique requiert donc des configurations matérielles lourdes afin d'utiliser les modèles de traduction tout en fournissant des performances normales.

Les modèles statistiques sont basés sur des corpus parallèles, c'est à dire que le même corpus est traduit dans une (ou plusieurs) autre(s) langue(s). Les alignements entre les textes peuvent être réalisés à différents niveaux : paragraphes, phrases, expressions et mots. Cependant, l'alignement mot-à-mot reste celui qui nous offre le plus d'information. Pour observer comment cet alignement est réalisé, veuillez vous référer au schéma 3.1 en annexe.

Un dictionnaire bilingue est alors construit à partir de ces bi-phrases et les probabilité de correspondances sont calculées. Le tableau ci-dessous représente ce dictionnaire dans exemple du français vers l'anglais.

la		the		0.6
la		this		0.4
femme		woman		0.5
femme		wife		0.5
la femme		the woman		0.6
a femme		this woman		0.4
de		of		0.6
ménage		household		0.8
femme de ménage		maid		1.0
a sauté		jumped		0.85
a sauté		skipped		0.15
un repas		a meal		1.0
sauté un repas		skipped a meal		1.0

FIGURE 2.2 – Dictionnaire bilingue avec probabilités

Chaque ligne du tableau représente une traduction et chaque élément de cette traduction est séparée par le symbole « ||| ». Dans la première colonne apparaît le mot ou la séquence de mots d'une langue source, dans la seconde colonne apparaît le mot ou la séquence de mots en langue cible. Puis, la dernière colonne représente la probabilité de correspondance de ce mot lors d'une traduction.

Dans la réalité, la dernière colonne contient une série de scores du modèle statistique. Mais nous avons choisi de n'afficher que le score du modèle de traduction afin de simplifier la compréhension. Le score du modèle de traduction est utilisé pour évaluer les différentes traductions possibles pour un même mot ou séquence de mot en langue source. Plus le score est proche de 1, plus la traduction a de chance d'être la bonne. Par exemple, la traduction de « la » en « the » a une probabilité de 0.6 (60%), c'est à dire que dans la majorité des cas, il faudra utiliser cette traduction et dans 40% des cas, on devra utiliser la traduction en « this ».

Lors d'une traduction, le système de traduction automatique statistique va commencer par segmenter le texte en mots, séquences de mots ou signes de ponctuations. Puis, il assemble un ensemble d'hypothèses de traduction utilisant les traductions que nous avons montrés dans la figure 3.1. Comme il existe de nombreuses façons de découper le texte et de traduire les séquences, la création des hypothèses produit une longue liste d'hypothèses de traduction. A chaque hypothèse on associe un score calculé à partir du modèle statistique. L'hypothèse ayant le plus gros score est alors choisi pour être la traduction que l'on va retourner.

Traduction par règles

La traduction à base de règle est un modèle de traduction se basant sur les informations linguistique de la langue source et de la langue cible de la traduction. Ces informations linguistiques sont récupéré à partir de dictionnaire bilingue, de règles de grammaire, de la sémantique, la morphologie et de la syntaxe de chaque langage. C'est à partir de ces données linguistiques qu'un système de traduction à base de règle pourra traduire des textes.

L'approche principale de ce système est de lier la structure de la phrase donnée avec la structure de la langue demandé, en maintenant le sens de la phrase.

Pour réalisé une traduction, ce système à donc besoin de :

- **Un dictionnaire** qui va lier chaque mot de la langue source avec les mots de la langue cible
- **Les règles** représentant les phrases communes de la langue source ainsi que de la langue cible
- **Des règles** liant les règles des langues entre elles (association des règles équivalentes)

Une fois toutes ces données rassemblées, voici les étapes de la traduction :

- **1.** Identifier chaque mot du texte source (nom commun, adverbe, verbe, article etc.)

- **2.** Recueillir les informations syntaxiques des verbes (base du verbe, temps, personne etc.)
- **3.** Analyse grammaticale de la phrase source (structure)
- **4.** Traduction des mots de la langue source en langue cible (en gardant le sens du mot, c’est à dire que l’on va traduire un adverbe avec un adverbe, un article avec un article etc. afin de garder le sens)
- **5.** Mise en correspondance des formes syntaxiques et production du résultat

La traduction à base de règles à un bon niveau de qualité pour des traductions dans un cadre général (hors domaines spécifiques). Mais elle nécessite un corpus de texte de référence (pour le dictionnaire) assez volumineux et elle demande de fait, une puissance de calcul relativement élevée.

Traduction par l’exemple

La traduction par l’exemple, également appelé traduction par analogie, réside dans le principe de prendre des petites phrases courtes qui sont traduites dans plusieurs langues et d’identifier ces phrases dans le texte à traduire. Cela permet une meilleure préservation du sens des phrases. Le résultat est donc plutôt satisfaisant.

La traduction par lexemple s’introduit entre la traduction par règles et la traduction statistique. En effet, beaucoup d’approches de la traduction intègrent des règles et des techniques statistiques. Cependant, certaines caractéristiques dissocient la traduction par l’exemple de la traduction par règle et de la traduction statistique. Dans la traduction par l’exemple, on isole le texte par phrase.

Voici les principales étapes de la traduction par l’exemple :

- **Décomposition** de la phrase en séquences correspondant le mieux aux exemples en base de données
- **Traduction** des séquences dans la langue cible par analogie avec les exemples en base de données

- **Recomposition** des séquences afin de former une phrase dans la langue d'origine

La traduction par l'exemple est particulièrement performante avec les verbes à particules - verbes qui changent de sens selon la particule (adverbe, adposition, nom ou adjectif) comme « mettre bas » ou « passer outre » en français - qui sont très présents dans les langues germaniques. Ce type de traduction est particulièrement performant avec ces verbes car il s'appuie sur des séquences de phrases, le contexte est donc conservé et la génération d'un verbe à particule est donc beaucoup plus aisée. Cependant, cette traduction nécessite une base de données très fournie pour fonctionner correctement. Cette base de données devra être alimentée par des textes de références traduits dans chaque langue. Il s'agit d'un lourd investissement.

2.2.3 Notre choix

Après avoir analysé les différents types de traductions automatiques, nous pouvons maintenant décider de quelle méthode de traduction nous allons nous servir. Dans le cadre de notre scénario, on souhaite ce qu'il y a de mieux, c'est pourquoi nous allons nous orienter vers le choix qu'on fait les leaders de ce marché : utiliser deux de ces règles. Nous allons donc utiliser à la fois la traduction par règle et la traduction statistique afin d'avoir un résultat le plus fiable possible.

En effet, les deux méthodes se complètent, la traduction automatique par règle est très bonne au niveau de la bonne formation des phrases, des verbes et autres, et la traduction automatique statistique est quand à elle performante pour des domaines de traductions spécifiques (le contexte, en sorte) et sa traduction d'expression ou mot à mot est plutôt performante également. De plus, ces deux méthodes nécessitent un corpus de texte bilingue volumineux, on pourra alors mutualiser ce corpus.

2.3 Synthèse vocale

2.3.1 Présentation et but

La synthèse vocale a vocation à émuler artificiellement la parole humaine à partir de données, dans notre cas un texte, tout en restituant au maximum une prosodie naturelle. De nos jours d'important progrès ont été faits quand à l'intelligibilité des voix de synthèse, le côté synthétique dû au hachage entre chaque syllabe ayant été adouci.

2.3.2 Principes générale

Nous nous intéressons donc ici plus particulièrement à la synthèse vocale à partir de texte (Text-To-Speech) qui permet donc la restitution de voix à partir d'un texte correcte d'un point de vue lexical. Aujourd'hui ces systèmes sont composés de deux modules principaux permettant de :

- **décomposer** le texte en un ensemble d'unités basiques qui seront ensuite interpréter par le synthétiseur
- **synthétiser** la voix à partir des informations produite par le premier module.

Traitement automatique du langage naturel

L'ensemble des modules permettant la décomposition du texte est désigner comme le module de traitement automatique du langage naturel. Plusieurs traitements consécutif sont alors exécuté sur le texte pour le transcrire sous une forme lisible par le synthétiseur vocale. On retrouve alors les traitements principaux suivant :

La phase de nettoyage

Le but est d'aplanir le texte en y remplaçant toutes les occurrences d'abréviations, de chiffre et nombre, de caractère spéciaux, unités de mesures ou encore devises par leur équivalent en toutes lettres transformant alors « num. tel. 01.00.00.00.01 \$2 la minute à partir de 08h30 » en « numéro téléphone zéro un, zéro zéro, zéro zéro, zéro zéro, zéro un deux dollars la minute à partir de huit heure trente ».

Dans les langues concernés, une étape de ré-accentuation peut être à prévoir par exemple dans les cas où les accents ne seraient pas présent de mots en majuscule. Lors de cette étape, il peut aussi être intéressant de repérer et traiter les acronymes comportant des voyelles qui ne sont pas lu comme un mot, et ainsi décomposer l'acronyme « PIB » est nécessaire alors que « OTAN » ou encore « PHP » peuvent être laisser comme tel.

L'étiquetage morphosyntaxique

Cela permettra ensuite de lever les ambiguïtés phonétiques sur la plupart des homographes. Cette étiquetage consiste à annoter chaque mots par son type (nom, verbes, adjectif, etc.) et sa forme (féminin, pluriel, infinitif, etc.).

Cette étiquetage permet alors par exemple de différencier la prononciation de « fier » dans la phrase « Il semble fier mais il ne faut pas s’y fier », où le fait de distinguer « fier » de l’adjectif masculin singulier de « fier » le verbe à l’infinitif nous permet de discerner sa prononciation. Cette étiquetage pourra être complété par une analyse contextuelle pour palier à certains cas où l’étiquetage ne serait pas suffisant comme dans le cas « Ses fils jouent avec des fils de fer » ou « fils » f-i-s et « fils » f-i-l ne sont tous deux des noms masculins pluriels, mais via une analyse contextuelle il est possible de les discerner dans le cas présent grâce à « de fer » qualifiant les « fils » f-i-l.

La phonétisation

Elle est la dernière étape constitutive de ce premier module. Elle permet alors à partir du texte nettoyé et étiqueté de produire en sortie un ensemble de marqueurs phonétiques qui seront ensuite interprétés par le synthétiseur. De nombreux formats existent pour produire un texte phonétisé, certains permettant d’y intégrer la prosodie qui sera nécessaire pour l’appui des émotions.

Synthétiseur vocale

Vient alors la retranscription orale des informations phonétiques (éventuellement accompagnée d’informations prosodiques) obtenue via le module de traitement automatique du langage naturel.

Dans un premier temps pour effectuer cette synthèse, les chercheurs se sont orientés vers des stratégies dites par règles, qui nécessitent très peu d’espace en mémoire (les sons ne sont pas enregistrés, ils sont générés via le traitement d’un signal). Cette approche donne malheureusement des résultats très éloignés de la voix humaine et nuit ainsi à la compréhension. Avec l’augmentation de la taille des espaces mémoire dans les systèmes informatiques une nouvelle approche a pu voir le jour : ici, plutôt que de générer entièrement le son, il est question de concaténer un ensemble de sons pré-enregistrés pour reproduire la parole. Cette deuxième approche, selon son implémentation,

permet alors des résultats bien plus proche de la voix humaine les sons émis n'étant que des enregistrements réels.

La synthèse par règles

Cette forme de synthèse aussi appelé synthèse par formants, sappuie alors sur la modélisation du spectre sonore de la parole en reproduisant généralement les trois premiers formants d'un phonème. Si cette technique a l'avantage d'être très légère, autant en mémoire qu'en traitement, elle ne permet que des résultats très éloignés de la voix humaine et dont la prosodie, et donc à fortiori la reproduction des émotions, n'est ni vraisemblable ni agréable. Ce processus n'est donc pas du tout adapté à nos besoins.

La synthèse par concaténation

Elle consiste en l'assemblage de plusieurs unités (pouvant varier) sonores près-enregistré pour reformer le texte à énoncer. Cette technique se basant alors sur des enregistrement de voix naturel elle permet des résultats bien plus naturels. Se pose alors deux questions :

- **L'unité** : en effet plus l'unité est longue (mot, phrases) moins de transition synthétiques sont à effectuer et plus le résultat final sera naturel. Malheureusement, cela implique de stocker énormément de données et donc, de devoir effectuer des traitement lourd lors de leur récupération. A contrario, un système se basant sur des unités courtes (phonème et diphones) plus le traitement sera simple et souple (possibilité d'intégrer plus facilement de la prosodie) mais plus il sera difficile de masquer les transitions entre celles-ci.
- **Le système de concaténation** : les systèmes les plus simples se contentant alors d'assembler bout à bout les unités, pour un rendu très peu naturel et haché. Pour répondre à ce problème il possible de simplement augmenter le nombre d'extraits correspondant à chaque unités ce qui permet ensuite de concaténer celles dont la transition est

la plus naturel et permet aussi d'émuler simplement les accentuation et marques prosodiques.

2.3.3 Dans notre cas

Dans notre système l'emphase au niveau de la synthétisation de la voix est à mettre du côté du naturel : nous visons en effet à rendre notre système capable de reproduire une émotion, il nous est donc crucial d'avoir un système capable de moduler le rythme, l'intonation et l'intensité de la voix de synthèse. Si notre système pourrait se rapprocher d'un système embarqué il est aussi vrai que les capacités de stockages récentes permette pour un prix faible de stocker plusieurs gigaoctets de données dans un volume très faible et nous ne considérerons donc pas les problèmes liée aux demandes de grand espaces de stockages. En revanche les capacité du système à produire une vocalisation vraisemblable est très importante, une voix à l'aspect trop robotique risquant d'affaiblir la restitution des émotions.

Si la synthèse par diphtonges permet des résultats intelligible et une flexibilité correcte, ceux-ci reste bien loin d'être naturel. Il en va de même pour toutes les solutions traitant un seul type d'unité. Nous nous orientons donc vers les systèmes à unité de longueur variable où chaque unité est représenté plusieurs fois pour permettre des transitions fluides entre chaque enregistrement.

Un tel système de synthèse vocale par concaténation se nomme alors synthèse par sélection d'unités. Ici la problématique tourne alors autour de la sélection des unités les plus pertinentes à piocher dans une base de donnée très conséquente (plusieurs dizaines d'heures d'enregistrement), il est donc question de sélectionner une unité correspondant déjà à nos besoin pour éviter d'avoir à la modifier ce qui résulte en un ensemble plus naturel. Pour la sélection des unité il est alors nécessaire d'essayer de minimiser deux paramètres : la différence entre l'unité à reproduire et l'unité à trouver ainsi que les différences entre la fin de l'unité précédente avec celle à trouver. Pour permettre un tel choix il est alors primordiale de disposer pour chaque unité de

variantes basé au minimum sur leurs durée et fréquences fondamentale (caractéristique acoustiques) ainsi que sur le ton (caractéristique symbolique, nécessaire aux émotions).

Ce type de traitement, si il nécessite une mise en place lourde reste le plus adapté à nos besoin en proposant la plus large variété de modulation de la prosodie sur le résultat tout en limitant les intervention sur le signal.

2.4 Émotions

2.4.1 Présentation et but

Dans cette partie, nous allons nous attarder sur les émotions. Mais plus précisément nous allons voir comment il est possible d'identifier une émotion, de quelle manière nous pouvons l'interpréter, l'utiliser et enfin, nous verrons comment il nous est possible de retranscrire ces émotions par le biais d'une synthèse vocale.

Mais avant de commencer, il nous semble important de bien définir ce qu'est une émotion. Une émotion est une expérience psychophysiologique complexe qui reflète l'état d'esprit d'un individu lorsqu'il est influencé par des facteurs internes ou externes. Il faut savoir que l'émotion est souvent associée à l'humeur, au tempérament, à la disposition et / ou à la motivation. Ce qui sera fort utile dans notre scénario puisque cette réunion sera professionnel, les personnes autour de la table pourront mesurer par la voix retransmise l'état d'esprit de la personne, sont humeur et surtout sa disposition et sa motivation à propos de l'objet de la réunion (un projet par exemple).

2.4.2 Identification des émotions

Quels sont les facteurs d'identification des émotions ?

L'identification des émotions, de manière générale, sont identifiables par 4 facteurs :

- Par une activité physique du corps
- Par une activité physique du visage
- Par des marques sémantiques d'émotions
- Par des marques phonologiques

L'activité physique du corps

Lorsque l'on ressent une émotion, il arrive que nous effectuons des gestes corporels conscient ou même inconscient. Cela peut venir d'un tique que nous avons face à une certaine émotion (même faible) ou alors cela peut être déclenché par une émotion forte.

Dans le cas où il s'agirait d'un tique, la personne va par exemple se ronger les ongles si elle commence à être stressé ou inquiète. Mais dans le même cas, une autre personne pourrait commencer à « jouer » avec ses doigts, les taper sur une table ou autre. En fait, ce genre de réaction face à une émotion dépend de chaque personne et peut varier selon les gens. Cependant, on peut quand même identifier certaines réactions, qui sont plutôt universelles dans le sens de ce qu'elle traduit (se ronger les ongles par exemple, tous le monde ne le fait pas mais lorsque quelqu'un le fait, c'est généralement parce qu'ils sont stressés).

En outre, dans le cas où il s'agirait d'une réaction provoquée par une émotion forte, cela se traduit généralement plus à un changement de comportement physique. Par exemple, en cas de fort énervement, il arrive qu'une personne commence à « gesticuler », qu'elle ne tienne plus en place. Ses mouvements deviennent plus rapides et plus fréquents. Ce genre de réactions sont plus généralisées chez les individus que la présence de tique, néanmoins toute

personne est différente et il peut arriver qu'une personne ne réagisse pas de la même façon face à une émotion forte.

L'activité physique du visage

Le visage d'un individu est l'élément physique qui transmet le plus d'émotions. En effet, le visage est la partie du corps qui exprime le plus les sentiments (et donc les émotions) d'un individu. De nombreuses études ont été menées afin d'identifier clairement les réactions faciales des individus face à une émotion. Nous maintenant chaque interprétations possible d'une émotion chez un humain.

Les marques sémantiques d'émotions

De manière générale, la structure des phrases d'un être humain selon ses émotions est assez figée, l'expression sémantique en revanche est extrêmement variée. On peut observer cette variété dans l'expression de demandes très semblables dans leur objet (même chez des personnes d'un même milieu de travail).

Par exemple, dans le cas d'une assistance technique au sein d'une entreprise, les personnes vont exprimer leur demande mais de façon différente en fonction de leurs émotions. Dans cet exemple, nous pouvons observer 4 catégories de marques sémantiques :

- **L'aspect technique** Par exemple : « Excel est inaccessible », « Les portables ne fonctionnent plus », « Je nai plus accès à mes adresses »
- **Les conséquences** Par exemple : « Je suis paralysée dans mon travail », « Jai un client qui simpatiente »
- **L'état psychologique** Par exemple : « Cest très pénible »
- **La demande** Par exemple : « Est-ce que vous pourriez nous arranger ça pour de bon ? »

Ces marques sémantiques sont très intéressantes car elles permettent de déterminer très rapidement l'état psychologique actuel de la personne. Et cela, rien qu'avec une phrase.

Les marques phonologiques

Les marques phonologiques ou plutôt données phonologiques sont les très importantes. Ce sont elles qui donneront le plus d'informations sur les émotions d'une personne grâce à la voix. Ce que l'on cherchera à analyser dans ce cadre sera le ton (calme, neutre, élevé) et le débit (lent, normal, rapide). Ces informations, combinées aux marques sémantiques, permettront alors de déterminer l'émotion d'une personne de manière fiable.

Afin de montrer comment nous pouvons analyser ces données phonologique et d'identifier précisément ce qu'elles sont, nous allons nous appuyer sur une étude du Laboratoire CLIPS (communication langagière et interaction Personne-Système) de Grenoble effectué par Solange Hollard, Mutsuko Tomokiyo et Denis Tuffelli intitulé « Une approche de l'expression oral des émotions : étude d'un corpus réel ».

Dans cette étude, ils analysent notamment les données phonologiques d'un énoncé qui est le suivant : « Jappelle pour deux problèmes, d'une part, donc nos deux ordinateurs sont euh ne peuvent pas être démarrés suite, suite à une coupure d'électricité cette nuit ».

Dans cet énoncé, ils compareront la durée, l'énergie et le pitch du mot numéraire « deux » qui est prononcé deux fois dans l'énoncé, la première de façon neutre et la deuxième de façon insistante.

Voici les mesures effectués sur les deux mots :

Mot prononcé	Énergie moyenne (en dB)	Durée (en secondes)	Pitch moyen (en hertz)	Mesures sur le mot précédent: énergie moyenne (en dB)	Mesures sur le mot précédent: pitch moyen (en hertz)	Écart avec le mot précédent: énergie moyenne (en dB)	Écart avec le mot précédent: pitch moyen (en hertz)
Deux (neutre)	77,3322	0,6672	250,1578	70,7792	232,1596	6,5530	17,9981
DEUX (avec émotion)	81,8968	0,8428	270,6249	73,4685	208,2545	8,4284	62,3704

FIGURE 2.3 – Mesures de durée, pitch et énergie sur un même mot prononcé de façon neutre, et avec émotion

On remarque alors que l'énergie, le pitch et même la durée sont supérieurs lorsque le mot est prononcé avec de l'émotion. On notera également l'écart avec l'énergie et le pitch du mot précédent qui est également plus élevé.

Grâce à cette analyse, on peut alors imaginer analyser chaque mot de la langue dans chaque émotion et ainsi se constituer une base de données contenant toutes les données phonologiques dont nous avons besoin afin d'identifier une émotion dans un énoncé d'une personne.

Comment allons-nous les identifier ?

Dans le cadre de notre scénario nous n'avons pas de caméra qui filme chaque personne de la réunion. Les expressions corporelles et du visage sont donc inutilisables. En revanche, nous pouvons utiliser les marques sémantiques décrites un peu plus haut ainsi que les données phonologiques que nous pourrions avoir grâce au micro dans lequel parleront les membres de la réunion.

Techniquement, nous allons donc récupérer l'enregistrement de la phrase prononcée par une personne puis nous allons analyser les données phonologiques, les comparer à notre base de données et nous allons donner à chaque

mot un score de probabilité pour chaque émotion. Ensuite, sur l'ensemble des mots de la phrase, nous déterminerons quelle émotion est la plus probable (en fonction des scores).

Parallèlement, nous analyserons les marques sémantiques de la phrase à traduire. Si les marques sémantiques peuvent sceller une émotion dans la phrase, on calculera la probabilité de l'exactitude que ce soit cette émotion là. Nous aurons déjà une probabilité pour l'émotion déterminée avec les données phonologiques, nous choisirons alors la probabilité la plus forte entre les deux système. Les probabilités pourront ne pas être équivalente, il pourrait arriver qu'une probabilité de 80% en données phonologiques soit supérieure en chance de réussite à une probabilité de 90% en marques sémantiques. C'est pourquoi, il faudra entrer notre système avec des humains qui vérifierons les résultats fournis par le système et qui ajusterons le calcul de la détermination de l'émotion choisi. Cette phase d'entraînement du système est très importante afin que le système soit juste et fiable.

2.4.3 Retranscription des émotions

L'étape finale de notre système est la retranscription orale de la traduction de la parole d'un locuteur. Dans le but de rendre cette traduction plus agréable et fidèle nous nous intéressons donc à la retranscription non seulement des dires mais aussi des émotions voisé par ce locuteur.

La reproduction des émotions dans la synthèse vocale est un élément de recherche qui n'est pas encore totalement résolu, pourtant les premières solutions datent du début des années 1990. C'est un sujet très complexe, en effet la reconnaissance d'une émotion uniquement par la voix n'est pas aisé et un humain pourra facilement confondre différentes émotions même lorsqu'elles sont produites par un acteur humain. Il reste d'ailleurs encore à prouver qu'une machine puisse parfaitement simuler n'importe quelle émotion de façon crédible et les recherches actuelles tendent à se concentrer sur un petit panel d'émotions facilement identifiable tel que la joie et l'énervement.

Un autre des freins aux recherches sur le sujet se trouve sûrement être le cadre très réduits des applications nécessitant réellement la reproduction des émotions dans une voix synthétique, ainsi la plus part des cas d'utilisations de voix de synthèse ne nécessite qu'un ton neutre comme pour les plates-formes d'aide téléphonique automatisé ou la lecture de contenu pour personnes mal voyantes. Il existerais néanmoins des domaines d'application tel que les voix de synthèses utilisé dans les prothèses à destination des personnes aux possibilités de communication orale réduite ou encore dans le cadre de notre scénario.

Néanmoins, maintenant que les progrès en terme de synthèse vocale permettent la restitution d'une voix neutre crédible les recherches s'orientent plus vers la reproduction des émotions, dernier pas vers des voix de synthèse réellement vraisemblables.

Retranscription par synthèse vocale

Systèmes à sélection d'unité

Ce système qui a déjà été présenté dans la section « Synthétiseur vocale » est le système le plus simple permettant de générer une voix capable d'émotion et aussi l'un des plus robuste et crédible : se basant entièrement sur le principe de modifier au minimum voir même de garder intacte les enregistrements du corpus la prosodie et la qualité de voix sont parfaites. Le plus gros désavantage de ce système se trouvant dans le travail d'enregistrement et d'annotation du corpus qu'il faut produire en amont.

Il est effectivement nécessaire d'enregistrer une quantité considérable de son pour parvenir à la constitution d'une base de donnée exhaustive et ce avec un seul locuteur pour toutes les émotions, aussi pour l'introduction d'une nouvelle émotion dans le système il sera nécessaire d'enregistrer à nouveau tout le contenu de la base avec cette nouvelle émotion. Il est possible de pallier le problème en proposant en plus des annotations symboliques manuels

une analyse des signes acoustique d'une émotion et les utiliser lors de la sélection, piochant ainsi dans les résultat proche en cas de trou. Si les systèmes à sélection d'unité son aujourd'hui les plus performant pour reproduire un éventail près-délimité d'émotion, et ce avec un résultat naturel ils sont aussi incapable de nuance : il faudrait enregistrer chaque nuance au préalable, ce qui n'est pas une solution praticable. Plus récemment un nouveau système, bien que moins efficace pour ma restitution d'une émotion appuyé, capable de nuance est demandant un travail d'enregistrement moindre a été développé.

Système basé sur les modèles de Markov caché

Ici la synthèse se fait via la concaténation d'une série de modèles de Markov caché dépendant du contexte qui détermineront la durée et les fondamentales du spectre puis la forme de l'onde est ensuite généré via un filtre dit de Mel Log Spectrum Approximation.

Ce système présente l'avantage d'être capable d'émuler plus finement des émotions en jouant sur des nuances. Cette approche présente l'avantage de nécessiter une base de donnée bien plus réduite.

Des recherches sur l'efficacité de ces deux méthodes ont été effectué. Aucune des deux techniques ne se sont réellement démarqué, en effet si les systèmes basé sur des modèles de Markov caché se sont montré plus efficace sur la prosodie et donc sur les émotions se basant principalement sur sa modification. En revanche, ils sont plus faible pour la reproduction d'émotions reposant sur des variations du spectre sonore. Il a aussi été noté que les systèmes à sélection d'unité permettaient une accentuation plus forte de l'émotion la rendant plus facilement reconnaissable. C'est d'ailleurs pour ce marquage plus fort de l'émotion que c'est cette dernière solution qui fût retenue.

Limites du système

Quelque soit le système employé les différents testes ont montré que les systèmes de synthèse vocales ne permettent pas à un écouteur de discerner plusieurs émotions aux signes vocaux proches (joie/surprise), de plus certaine émotion tel que le dégoût on été très mal reconnues par les écouteurs humains. Ces deux systèmes sont donc encore sujet à améliorations.

Chapitre 3

Aller plus loin

3.0.4 Au niveau de la détection des émotions

Il serait intéressant de rendre le système encore plus fiable et plus compétent. Pour ce faire on pourrait intégrer un système de caméras au centre de la table de réunion qui filmerai les visages des participants afin d'en déterminer les émotions grâce à une banque de faciès riche. L'analyse de visage et d'émotions sur les visages étant un domaine complexe et attrayant, il serait intéressant de l'étudier plus en détails et l'utiliser dans notre scénario.

Si l'on souhaite pousser la détection des émotions à un plus haut niveau, on peut intégrer dans notre scénario un ensemble de capteurs qui mesurerais l'activité cardiaque de chaque personne car en fonction des émotions de base, le rythme cardiaque d'un individu évolue. Cette approche serait très pertinente et permettrait une amélioration des résultats significatives. Mais cette approche n'est pas totalement surréaliste car il existe aujourd'hui des capteurs cardiaques relativement précis et discret (bracelet au poignet). Cela ne provoquerai ni gêne, ni encombrement pour les utilisateurs.

3.0.5 Émotions et traduction

Un deuxième axe d'amélioration se situe au niveau de la traduction, et plus particulièrement de la prise en compte des données émotionnelle dans celle-ci. Il serait ainsi intéressant de se pencher sur la sélection du vocabulaire (niveau, nuance) à employer lors de la traduction en incorporant dans les paramètres de cette dernière l'émotion détecté lors de la prononciation de la phrase. Cela permettrait d'appuyer sur l'état d'esprit du locuteur de façon naturel via l'intégration d'un lexique contextualisé.

3.0.6 Synthèse de l'émotion

Nous l'avons vu, si les techniques actuels de synthèse vocale sont maintenant capable de reproduire une voix presque naturel lors de l'émulation d'une voix neutre il en est tout autrement lors ce que l'on tente d'incorporer des émotions dans le signal. Si la technique actuel ne permet pas encore aux modèles de Markov caché de rivaliser avec les systèmes à sélection d'unité les plus perfectionné ils sont bien l'avenir de la synthèse vocale. En effet, ces dernier on l'avantage de pouvoir donner des résultat correcte à partir de corpus modeste et sont finalement plus performant lors ce qu'il s'agit produire une voix neutre notamment grâce à leur capacité à suivre une prosodie plus naturel. De plus, le plus gros point faible des système à modèle de Markov caché réside dans la qualité des voix générés qui sont dégradé, néanmoins de nouvelles solutions alternative à celle exposé ici sont actuellement en développement, certaine tentant de reproduire le système de modulation de la voix chez l'humain.

Nous sommes tout de même encore loin de reproduire le langage naturel via un synthétiseur vocale, mais les avancés qui on été effectué sur les voix neutres permettant aujourd'hui de produire des voix satisfaisante à l'écoute, l'étape suivante tend invariablement à l'amélioration de la simulation des émotions.

Annexes

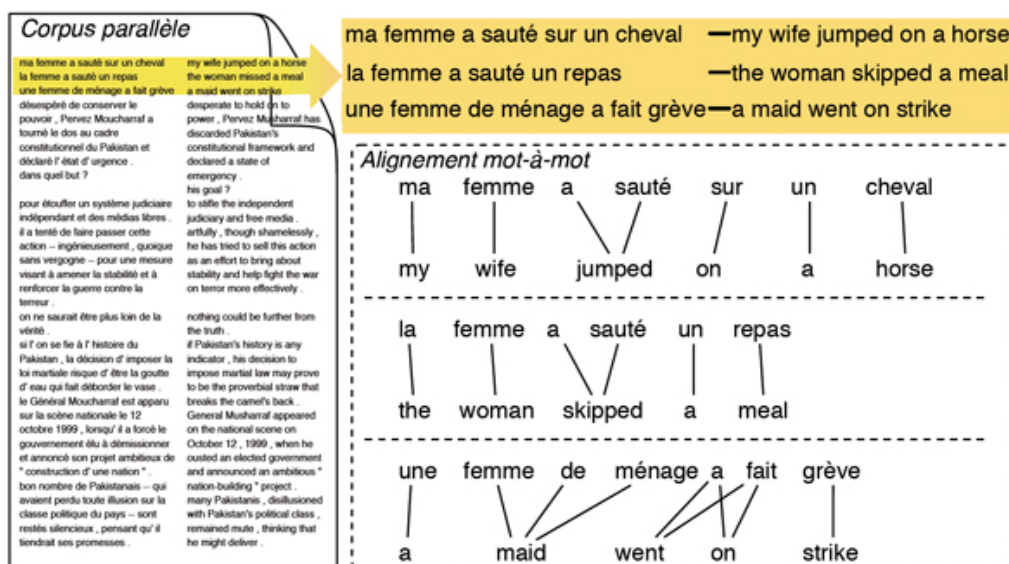


FIGURE 3.1 – Corpus Parallèle

Références

- **Appariement de phrases courtes pour la traduction automatique par lexemple** par Julien Gosme (2009) - Utilisation dans la partie "Traduction par l'exemple"
- **Traduction automatique** par Wikipédia - Utilisation dans la partie "Traduction"
- **Traduction** par Wikipédia - Utilisation dans la partie "Traduction"
- **Introduction à la traduction guidée par lexemple (Traduction par analogie)** par Michael Carl (2003) - Utilisation dans la partie "Traduction par l'exemple"
- **Comprendre la Traduction Automatique** par Systran - Utilisation dans la partie "Traduction statistique" et "Traduction par règles"
- **La traduction automatique statistique, comment ça marche ?** par Li Gong (2013) - Utilisation dans la partie "Traduction statistique"
- **Rule-based machine translation** par Wikipédia (EN) - Utilisation dans la partie "Traduction par règles"
- **Émotion** par Wikipédia - Utilisation dans la partie "Émotions"
- **Computatione Affective : Affichage, Reconnaissance, et Synthèse par Ordinateur des Émotions** par Marco Paleari (2009) - Utilisation des la partie "Émotions"
- **Une approche de l'expression orale des émotions : étude d'un corpus réel** par Solange Hollard, Mutsuko Tomokiyo et Denis Tuffelli (2005) - Utilisation dans la partie "Émotions"
- **Reconnaissance automatique de la parole guidée par des transcriptions a priori** par Benjamin LECOUTEUX (2008) - Utilisation dans la partie "Reconnaissance automatique de la parole"

- **Reconnaissance automatique de la parole** par Jean-Paul HATON (2013) - Utilisation dans la partie "Reconnaissance automatique de la parole"
- **Emotional Speech Synthesis** par Gregor O.Hofer (2004) - Utilisation dans la partie "Synthèse vocale"
- **STATISTICAL PARAMETRIC SPEECH SYNTHESIS** par Alan W Black, Heiga Zen et Keiichi Tokuda (2007) - Utilisation dans la partie "Synthèse vocale"
- **HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering** par Tuomo Raitio, Antti Suni, Junichi Yamagishi, Hannu Pulakka, Jani Nurminen, Martti Vainio et Paavo Alk (2011) - Utilisation dans la partie "Émotions"
- **Emotional Speech Synthesis : A Review** par Marc Schröder (?) - Utilisation dans la partie "Émotions"
- **Synthèse vocale par sélection d'unité : une méthode pour la redéfinition de la courbe intonative** par Baris Bozkurt, Thierry Dutoit and Vincent Pagel (2002) - Utilisation dans la partie "Émotions"
- **Analysis of Statistical Parametric and Unit Selection Speech Synthesis Systems Applied to Emotional Speech** par Roberto Barra-Chicote, Junichi Yamagishi, Simon King, Juan Manuel Montero, Javier Macias-Guarasa (2009) - Utilisation dans la partie "Synthèse vocale"