Mémoire

Florian Constant et Aubry Gudel

23 février 2014

Table des matières

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 dune table afin davoir 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 puisquil traduira en temps réel, cest-à-dire quil fera la traduction alors même que linterlocuteur na pas fini sa phrase. De plus, ce système sera capable danalyser 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 discutions pour notre système. En effet, que le système sache dans quelle contexte se situe la discussion est très important car il permettra daméliorer 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. Cest pourquoi nous allons partir du principe que quelquun 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 didentifier chaque personne si le système navait quun micro central. Cela fonctionnerait certainement mais dans ce cas, comment restituer dans la langue de la personne la traduction sans avoir une cacophonique de traduction dans la salle ? Il nous paraît alors évident quil soit nécessaire que chaque participant ai une oreillette dans laquelle on lui traduit les discussions.

Afin déviter un processus didentification 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 linterlocuteur. De cette manière, chaque personne sera identifiée de manière simple et sûre.

Pour ce qui est de la seconde question, cest-à-dire de comment connaître la langue, il sagira de la détecter en fonction de ce que la personne dira. Un peu à la manière dun Google Translate avec lécrit. Mais nous reviendront sur la réalisation de cette tâche plus tard.

Reconnaissance automatique de la parole A ce stade, il nous faut encore réussir à mettre sous format textuel ce quune personne dit afin de le traduire par la suite. Il existe des logiciels spécialisé 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, cest pourquoi nous nallons pas plus en parler pour le moment et estimé que ce travail est fait.

Traduction Maintenant que lon 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 dexprimer 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 vas 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 synthétisation 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.

Chapitre 2

Résolution du système

2.1 Reconnaissance automatique de la parole

Introduction La reconnaissance automatique de la parole (que nous abrégerons 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 était limité à la reconnaissance de mots isolé 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 retranscrire 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 des fenêtrant le signal d'origine et 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 permets 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 liée à lenvironnement (réverbération de la salle, bruit ambiant). Nous nétudierons pas ici les différentes solutions qui existent permettant damoindrir l'impacte des deux facteurs précédemment énoncé et passerons directement à l'étude du signale 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 fournier puis en analysant le résultat avec un ensemble de filtres passe-bande permettant le représentation de léchantillon 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 tel 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 temporel. Pour ce faire il existe plusieurs modèles notamment les modèles Markoviens Cachées, les modèles neuronimétriques ou encore les algorithme de comparaison dynamique.

La technique la plus couramment utilisé 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 isolent 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 noeuds 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 synthétisation 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 adapté 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 légereté du dictionnaire comportant les mots reconnaissable. Nous utiliserons donc le

phonème. Les autres paramètres nayant pas dimpact particulière sur le cur de notre recherche nous sélectionnerons le système le plus utilisé, les Modèles de Markov cachés adjoint à , ceux là ayant déjà prouvé leurs efficacités.

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.

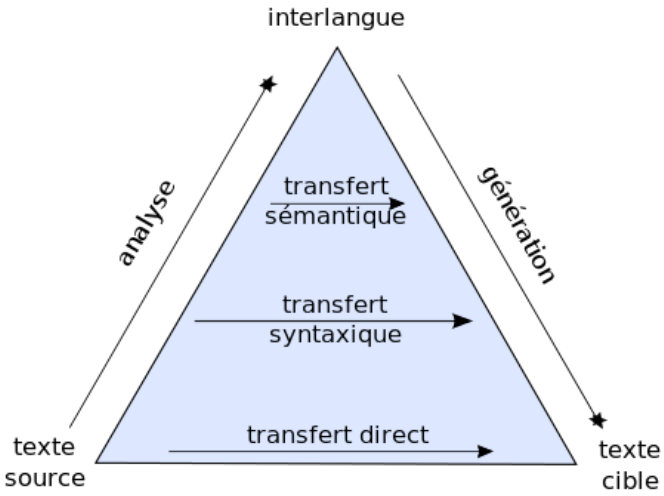

FIGURE 2.1 – Triangle de Vauquois

Ce schéma est représentatif des différents chemin 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 4.1 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 4.1. 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 élevé.

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 dapproches 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 présent dans les langues germaniques. Ce type de traduction est particulièrement performant avec ces verbes car il sappuie sur des séquences de phrases, le contexte est donc conservé et la génération d'un verbe à particule est donc beaucoup plus aisé. Cependant, cette traduction nécessite une base de données très fournie pour fonctionner correctement. Cette base de données devra être alimenté par des textes de références traduit dans chaque langue. Il s'agit d'un lourd investissement.

2.2.3 Notre choix

Après avoir analyser 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 orientée 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 naturel. De nos jours d'important progrès ont été fait quand à l'intelligibilité des voix de synthèse, le côté synthétique dû au hachage entre chaque syllabe ayant été adoucie.

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 lexicale. 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 dont le but est daplanir 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 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 permets 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 permets de discerner sa prononciation. Cette étiquetage pourra être compléter par une analyse contextuel pour palier à certain 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 contextuel il possible de les discerner dans le cas présent grâce à "‘de fer"’ qualifiant les "‘fils"’ f-i-l.

La phonétisation est la dernière étape constituante 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 interprété par le synthétiseur. De nombreux formats existent pour produire un texte phonétisé, certain permettant d'y intégrer la prosodie qui sera nécessaire pour l'appuie des émotions.

Synthétiseur vocale

Vient alors la retranscription oral des informations phonétiques (éventuellement accompagné d'informations prosodiques) obtenue via le module de traitement automatique du langage naturel. Dans un premier temps pour effectuer cette synthétisation les chercheurs se sont orienté vers des stratégies dites par règles, qui nécessite très peux d'espace en mémoire (les sons

ne sont pas enregistré, ils sont générés via le traitement d'un signale). Cette approche donne malheureusement des résultats très éloigné de la voix humaine et nuis ainsi à la compréhension. Avec l'augmentation de la taille des espaces mémoires 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 près-enregistré 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 , 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 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 (phonéme 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 synthètisation 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 diphones 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 certaines émotion (même faible) ou alors cela peut être déclenché par une émotion forte.

Dans le cas où il sagirait d'un tique, la personne va par exemple se ronger les ongles si elle commence a 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 a une émotion dépend de chaque personne et peut varier selon les gens. Cependant, on peut quand même identifier certaines réaction, qui sont plutôt universelle dans le sens de ce qu'elle traduise (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 provoqué 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 généraliser 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 menés 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, lexpression 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 a 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 appuyé 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 lexpression 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, dune part, donc nos deux ordinateurs sont euh ne peuvent pas être démarrés suite, suite à une coupure délectricité cette nuit ».

Dans cette 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 lénergie, le pitch et même la durée sont supérieurs lorsque le mot est prononcé avec de l'émotion. On notera également lécart avec lénergie et le pitch du mot précédent qui est également plus élevé.

Grâce a 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 film 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 pourront avoir grâce au micro dans lequel parleront les membres de la réunion.

Techniquement, nous allons donc récupérer l'enregistrement de la phrase prononcé 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 desceller 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 dentraînement du système est très importante afin que le système soit juste et fiable.

Les limites

2.4.3 Retranscription des émotions

Travaux sur le sujet

Influence sur la traduction

Retranscription par synthèse vocale

Limites du système

Chapitre 3

Aller plus loin

TODO

Chapitre 4

Annexes

FIGURE 4.1 – Corpus Parallèle