

Introduction

L'utilisation de connaissances supplémentaires conjointement au signal de parole est une méthode classique pour améliorer les performances et la robustesse des systèmes de reconnaissance automatique de la parole. De nombreux travaux sur la perception de la parole ayant montré l'importance des informations visuelles dans le processus de reconnaissance chez l'homme, l'utilisation de données sur la forme et le mouvement des lèvres du locuteur semble être une voie prometteuse pour la reconnaissance automatique surtout en milieux sonores bruités.

Les êtres humains emploient l'information visuelle de façon subconsciente afin de comprendre les paroles, particulièrement dans des environnements bruyants, mais également quand les conditions acoustiques sont bonnes. Le mouvement des lèvres du locuteur apporte une série d'information importante, par exemple au sujet des articulations, ce qui est automatiquement intégré par le cerveau. L'effet McGurk (1976) en apporte la preuve en montrant que le cerveau, soumis à des stimuli auditifs et visuels inconsistants, perçoit un son différent de celui qui a été dit.

1. Contexte et cadre de recherche

L'objet de nos travaux de recherche concerne l'intégration des informations visuelles aux informations acoustiques en vue de leur exploitation pour la reconnaissance automatique de la parole. Si cette exploitation est fort séduisante, la problématique qu'elle soulève est cependant loin d'être simple. Tout d'abord, se pose la question du niveau d'intégration : est-ce le niveau des données ou bien celui des résultats. Puis il y a les phénomènes de décalage temporel entre la réalisation auditive et la réalisation visuelle d'un même phonème. Ensuite intervient le problème d'adaptation des contributions des modalités acoustique et visuelle selon leur fiabilité relative. Enfin se pose la question de la pertinence de l'utilisation, pour le traitement du signal visuel de parole, d'unités de décision spécifiques, nommées visèmes.

Reconnaissance automatique de la parole audio-visuelle (AVASR) a été lancée par Petajan (1984) et elle est encore une zone active de recherche. Cette thèse se positionne clairement dans le champ des systèmes AVASR, le système de reconnaissance proposé utilise les modèles de Markov cachés (Hidden Markov Model, HMM) comme moteur de reconnaissance combiné avec un algorithme génétique (Genetic Algorithm, GA) pour résoudre le problème de la convergence vers l'optimum local. Le point principal de ce travail

est basé sur la qualité de la modélisation des données (appelé observations) faites par HMM. Notre objectif est de proposer des algorithmes qui permettent d'améliorer cette qualité. Le critère utilisé pour quantifier la qualité de HMM est la probabilité qu'un modèle donné génère une observation donnée. Pour résoudre ce problème, nous utilisons comme nous l'avons déjà mentionné une hybridation génétique des HMM et nous proposons des méthodes de représentation d'un gène et la méthode pour l'évaluation des mesures de remise en forme des populations de chaque génération crée par algorithme génétique. L'expérience est menée afin d'évaluer chaque population et la précision de résultat d'inférence sur un ensemble de données audiovisuelles réelles.

Le traitement de la parole arabe est encore à ses débuts, la raison pour laquelle, nous avons pensé à l'application de la méthode hybride GA/HMM, ayant comme objectif la reconnaissance de la parole en mode multi-locuteur.

2. Plan de la thèse

Ce document est structuré en deux parties. La première partie établit plusieurs états de l'art sur les domaines abordés (chapitre 1, 2 et 3) tandis que la deuxième partie présente nos approches proposées.

Le premier chapitre donne une brève présentation de quelques éléments physiologiques sur la production de la parole et la paramétrisation des lèvres. Cette étude présente une description des muscles faciaux intervenant dans le processus de la parole. Nous nous décrivons aussi quelques propriétés intrinsèques de la perception de la parole bimodale afin de mieux comprendre ce processus différents modèles d'intégration audiovisuelle chez l'homme et dans la machine sont présentées.

Nous passons dans le chapitre 2 à une description détaillée des techniques d'extraction des informations visuelles des mouvements des lèvres, notamment celles basées sur le traitement vidéo, ainsi que notre méthode de calcul des paramètres labiaux basée sur un maquillage préalable des lèvres.

Par la suite dans le chapitre 3, Nous nous consacrons a une revue de l'état de l'art dans le domaine du développement des systèmes de reconnaissance visuelle et audiovisuelle.

Chapitre 4 définit le principe et le fonctionnement de notre système AVASR proposé en utilisant la méthode hybride GA/HMM.

Puis dans le chapitre 5, nous décrivons le principe et le fonctionnement de notre système de reconnaissance de la parole audiovisuelle proposé.

Le dernier chapitre (chapitre 6) présente les résultats de nos tests sur les deux modèles d'intégration couramment utilisés dans la littérature (précoce et tardive) en insistant sur notre architecture d'intégration originale, basée sur une pondération des canaux en fonction de leur fiabilité, estimée par la dispersion des meilleurs candidats. La dernière partie de ce manuscrit est dédiée à la description technique de notre système électronique d'extraction des paramètres labiaux en temps réel et à l'évaluation de ses performances dans une application de lecture labiale automatique.

Enfin, nous concluons par un bilan de nos travaux de recherche et nous proposons quelques perspectives d'amélioration associées aux différentes réalisations.

Les lèvres et la production de la parole

1

Le téléphone et la radio prouvent la capacité d'une parole purement auditive à transmettre avec efficacité une communication langagière. Néanmoins, la perception humaine tire aussi profit de l'information visuelle apportée par le visage du locuteur notamment lorsque les conditions acoustiques sont dégradées. C'est cette bimodalité intrinsèque, et le gain d'intelligibilité qu'elle apporte, qu'explore l'étude de la parole audiovisuelle. Mise en évidence pour la communication humaine, elle ouvre de nouvelles perspectives pour la communication avec et par la machine.

Bien que la communication orale engage l'ensemble du visage du locuteur, les lèvres occupent une place privilégiée : elles fournissent une source visuelle d'information pour la perception de la parole et, étant toujours identifiables, se prêtent à une analyse automatique. La capture automatique des mouvements labiaux (ou la biométrie) tend à doter l'ordinateur de paramètres intelligibles et indépendants pour contrôler des visages synthétiques parlants ou bien identifier le message énoncé par une reconnaissance audiovisuelle automatique. Les difficultés technologiques résident dans la complexité de ces mouvements et la variabilité intra- et inter- locuteurs.

1.1 Architecture et fonctionnement de l'appareil vocal

Cette section, qui rappelle l'architecture et les principes généraux de fonctionnement de notre appareil vocal s'appuie sur les ouvrages suivants : (Le Huche 2001) et (Boite *et al.* 2000). Une vue schématique de notre appareil vocal est proposée à la figure 1.1.

1.1.1 L'appareil vibreur

L'air est la matière première de la voix. Si le fonctionnement de notre appareil vocal est souvent comparé à celui d'un instrument de musique, il doit être décrit comme celui d'un instrument à vent. En effet, en expulsant l'air pulmonaire à travers la trachée, le système respiratoire joue le rôle d'une soufflerie. Il s'agit du « souffle phonatoire » produit, soit par l'abaissement de la cage thoracique, soit dans le cadre de la projection vocale par l'action des muscles abdominaux.

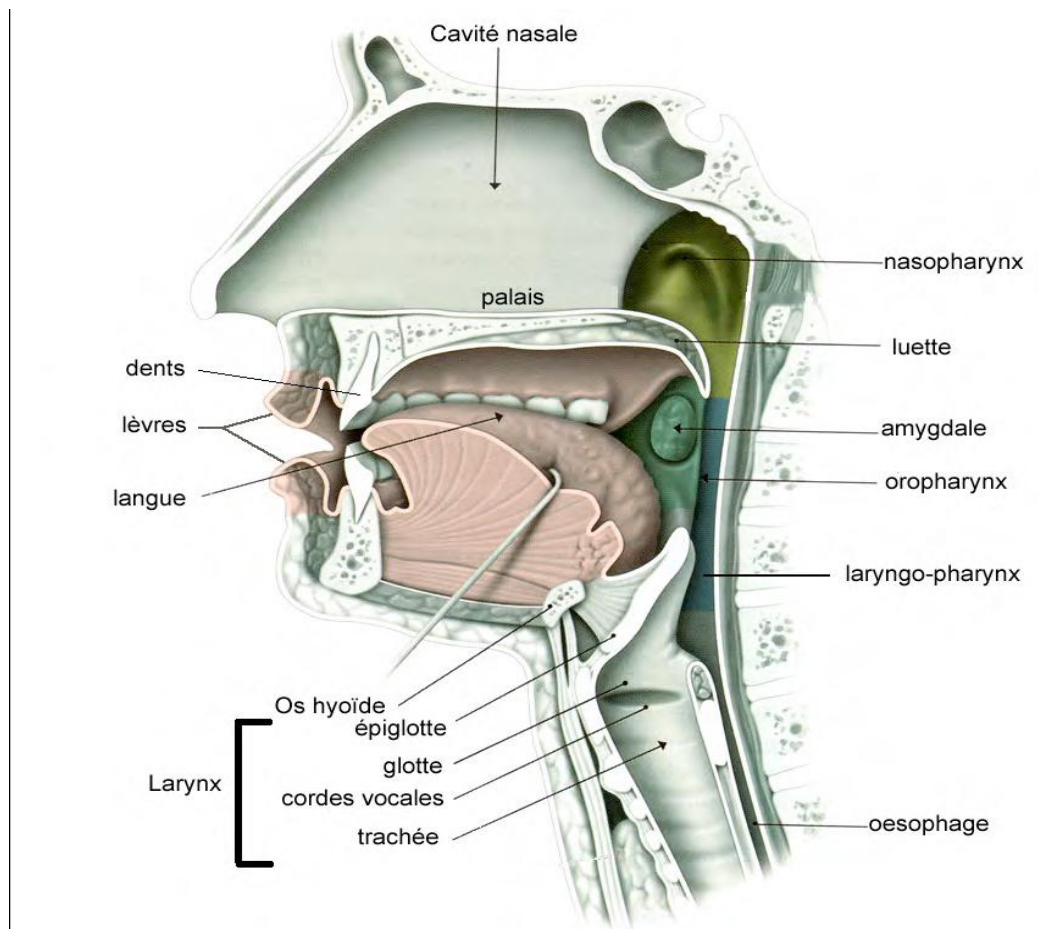


Figure 1.1 – Vue schématique de l'appareil vocal, dans le plan sagittal médian.

L'extrémité supérieure de la trachée est entourée par un ensemble de muscles et de cartilages mobiles qui constituent le larynx. Le plus important est le cartilage thyroïde qui forme le relief de la pomme d'Adam. Le larynx se trouve au carrefour des voies aériennes et digestives, entre le pharynx et la trachée, et en avant de l'œsophage. Les plis vocaux, communément nommés « cordes vocales » sont deux lèvres symétriques (structures fibreuses) placées en travers du larynx. Ces lèvres se rejoignent en avant et sont plus au moins écartées l'une de l'autre sur leur partie arrière (structure en forme de V); l'ouverture triangulaire résultante est nommée glotte. Les structures du larynx et des plis vocaux sont illustrés à la figure 1.2. Le larynx et les plis vocaux forment notre « appareil vibreur ».

Lors de la production d'un son qualifié de « non-voisé » (ou sourd), comme c'est le cas, par exemple, pour les phonèmes [s] ou [f], les plis vocaux sont écartés et l'air pulmonaire circule librement en direction des structures en aval.

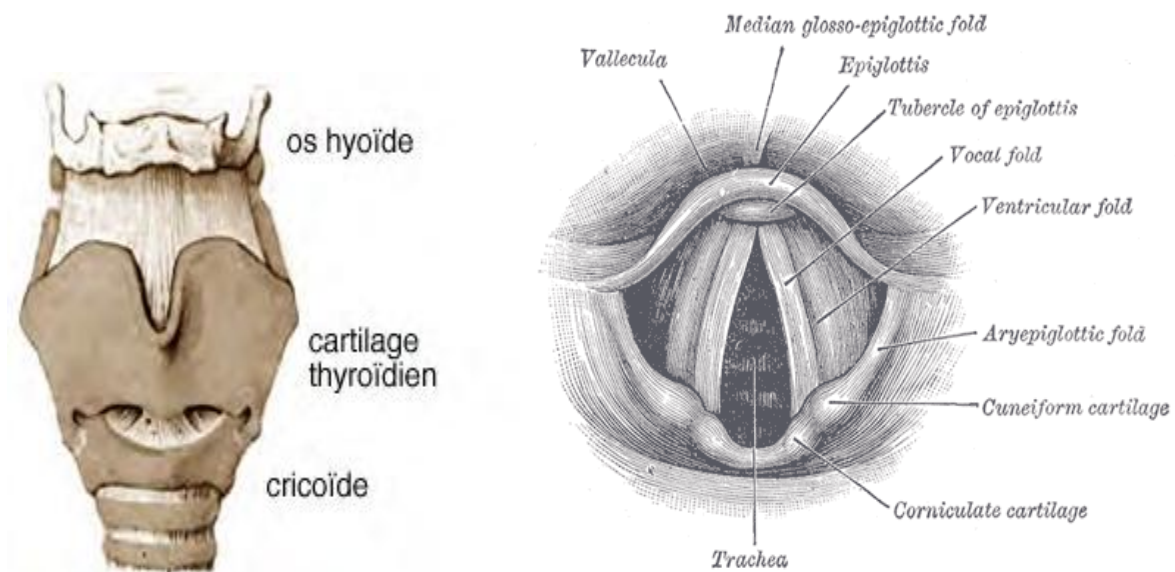


Figure 1.2 – Vue schématique antérieure du larynx (à gauche). Vue laryngoscopique des cordes vocales (à droite).

En revanche, lors de la production d'un son voisé (ou sonore), comme c'est le cas, par exemple, pour les phonèmes [z], [v] et pour les voyelles, les plis vocaux s'ouvrent et se ferment périodiquement, obstruant puis libérant par intermittence le passage de l'air dans le larynx. Le flux continu d'air pulmonaire prend ainsi la forme d'un train d'impulsions de pression ; nos « cordes vocales vibrent ». Le dernier élément principal de notre appareil vibreur est l'épiglotte. Lors de la déglutition, cette dernière agit comme un clapet qui se rabat sur le larynx, conduisant les aliments vers l'œsophage en empêchant leur passage dans la trachée et les poumons (« fausse route »).

1.1.2 Le résonateur

L'air pulmonaire, ainsi modulé par l'appareil vibreur, est ensuite appliqué à l'entrée du conduit vocal. Ce dernier est principalement constitué des cavités pharyngiennes (laryngopharynx et oropharynx situés en arrière-gorge) et de la cavité buccale (espace qui s'étend du larynx jusqu'aux lèvres). Pour la réalisation de certains phonèmes, le voile du palais (le velum) et la luette qui s'y rattache, s'abaissent, permettant ainsi le passage de l'air dans les cavités nasales (fosses nasales et rinopharynx ou nasopharynx). Ces différentes cavités forment un ensemble que nous qualifierons ici de « résonateur ». Si l'appareil vibreur peut être décrit comme le lieu de production de « la voix », le résonateur apparaît alors comme le lieu de naissance de « la parole ». Il abrite en effet des organes mobiles,

nommés articulateurs, qui en modifiant sa géométrie et donc ses propriétés acoustiques, mettent en forme le son laryngé (ou son glottique) en une séquence de sons élémentaires. Ces derniers peuvent être interprétés comme la réalisation acoustique d'une série de phonèmes, unités linguistiques élémentaires propres à une langue. Les articulateurs principaux sont la langue, les lèvres, le voile du palais et la mâchoire (maxillaire inférieur).

L'articulateur principal de la cavité buccale est la langue. Intervenant dans la mastication et la déglutition, la langue est également l'organe du goût. S'étendant sur une longueur d'une dizaine de centimètres environ, cet organe complexe et hautement vascularisé est composé d'un squelette, de muscles et d'une muqueuse. Son squelette est qualifié d'ostéofibreux ; il est constitué de l'os hyoïde, situé au dessus du larynx, sur lequel se fixe la membrane hyoglossienne, d'une hauteur d'un centimètre environ, et le septum lingual, lame fibreuse à l'origine de la dépression visible sur toute la longueur de la langue. Son mouvement est contrôlé par dix sept muscles, dont huit paires de muscles agonistes/antagonistes. Quatre paires de muscles extrinsèques (muscles qui prennent naissance à l'extérieur de la langue) servent notamment à sa protrusion, sa rétraction, sa dépression ou son élévation.

La langue est usuellement décrite comme un ensemble de deux structures au comportement distinct, la racine (ou base), fixée à l'os hyoïde, et le corps, plus mobile. Ce dernier se décompose également en deux parties, le dos et la pointe de la langue, nommée apex. L'organisation du système musculaire de la langue ainsi que ses principales structures sont illustrées à la figure 1.3. Le rôle de la langue dans la phonation est déterminant, notamment pour la production des voyelles, caractérisée par le libre passage de l'air dans le résonateur. La phonétique articulatoire décrit le système vocalique d'une langue (classification des voyelles) précisément à l'aide de deux critères qui décrivent la configuration de la langue dans la cavité buccale. Le premier est le « lieu d'articulation » ; « avant » ou « arrière », il localise la masse de la langue et qualifie ainsi les voyelles produites d'« antérieures », de « centrales » ou de « postérieures ». Le second critère est « l'aperture » ; il décrit l'espace de résonance ménagé entre la langue et le palais (fermé ou ouvert), qualifiant ainsi les voyelles produites de « hautes » ou « basses ». La langue joue également un rôle important pour l'articulation des consonnes, dont le mode de production est, à l'inverse des voyelles, caractérisé par l'obstruction du passage de l'air dans le résonateur. Dans ce cas, le « lieu d'articulation » localise cette obstruction. Pour produire une consonne dite « dentale » ([t], [d], [n]), la pointe de la langue crée cette obstruction en se rapprochant des dents.

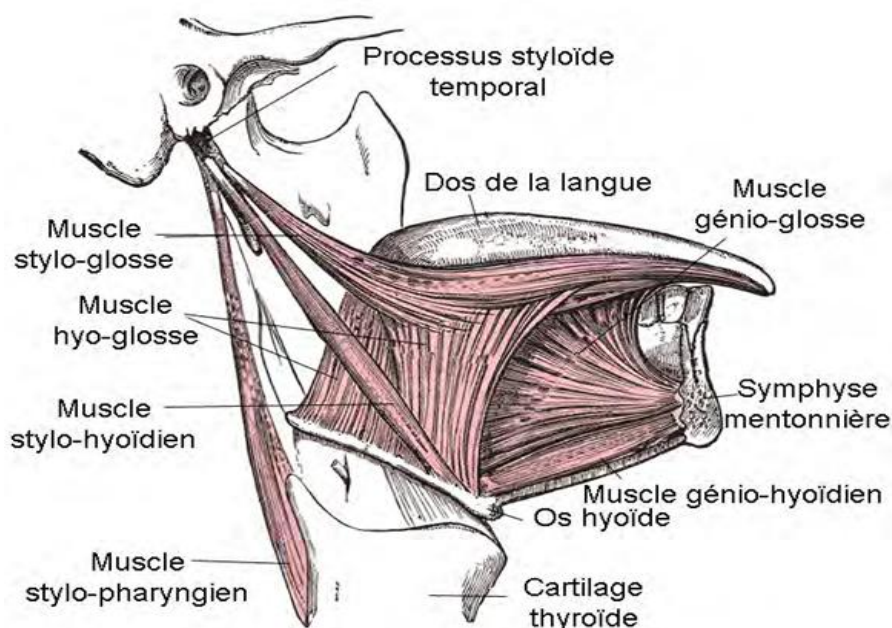


Figure 01.3 – Structures de la langue, détails des muscles extrinsèques (plan sagittal médian, vue de droite).

Dans le cas des consonnes « inter-dentales » ([th] comme *thin*, [dh] comme *then*), la langue dépasse les dents et vient s'appuyer directement sur les incisives. Pour les consonnes dites « alvéolaires » ([s], [z] ou la consonne liquide [l] mais également [t], [d], [n]), et « postalvéolaires » ([ch] comme *church*, [jh] comme *judge*, [sh] comme *she*, [zh] comme *azure*), elle se déplace respectivement vers les alvéoles (creux de l'os alvéolaire dans lequel est enchâssée une dent) et vers la partie antérieure du palais (à la juxtaposition avec le palais dur). Pour une consonne dite « palatale » ([j] comme *ye*, catégorisée également comme une semi-voyelle), l'organe articulateur est le dos de la langue, l'obstruction ayant lieu au niveau du palais dur. Pour une consonne vélaire ([k], [g], [ŋ] comme *parking*), la partie postérieure du dos de la langue se bombe et se rapproche du palais mou. Enfin, pour une consonne uvulaire ([r] comme *Paris* en français), le lieu d'articulation se situe au niveau de la luette.

Les lèvres constituent l'autre articulateur majeur de la cavité buccale. Elles permettent la production des consonnes « bilabiales » (rapprochement des lèvres inférieures et supérieures, [p], [b], [m]) et des consonnes « labio-dentales » ([f], [v], rapprochement de la lèvre inférieure avec les dents). Elles interviennent également dans le cadre de la production vocalique en apportant la notion d'arrondissement des voyelles. Enfin, la réalisation acoustique de certains phonèmes nécessite parfois deux lieux d'articulation, impliquant à la fois la langue et les lèvres ; c'est le cas notamment de la consonne « labio-velaire » [w] (comme *who*).

Le dernier articulateur du résonateur est le voile du palais qui permet, lorsqu'il s'abaisse, de mettre en parallèle les cavités buccale et nasale. Il intervient notamment dans la production des consonnes nasales [m], [n] et [ng] en les différenciant respectivement des groupes de consonnes ([p], [b]), ([t], [d]), et ([k], [g]), qui présentent la même configuration linguale et labiale. Enfin, l'abaissement du voile du palais permet, en langue française notamment, la formation des voyelles nasales [ɔ] (*on*), [ɛ] (*hein*), [œ] (*un*), [ɑ] (*an*).

Au regard de ces principaux résultats issus de la phonétique articulatoire, la réalisation acoustique d'un phonème dépend principalement des configurations de la langue, des lèvres et du voile du palais mais également de l'activité des cordes vocales. Lorsque ces dernières doivent être retirées, dans le cadre notamment du traitement chirurgical du cancer du larynx, les mécanismes de la phonation sont profondément modifiés.

1.2 L'anatomie des lèvres

1.2.1 Les tissus

D'après les données anatomiques présentées dans (Abry 1980), les lèvres forment deux replis musculaires, recouverts d'une membrane, qui circonscrivent l'orifice de la cavité buccale. Ces replis supérieur et inférieur sont indépendants et se réunissent à leurs extrémités pour former les commissures labiales. La face externe des lèvres est recouverte par de la peau et la face interne par de la muqueuse composée de cellules disposées comme des pavés (l'épithélium). Les muscles se trouvent directement sous la peau.

La ligne entre la peau et la muqueuse dessine dans sa partie supérieure et, au centre, une courbe concave dénommée « arc de Cupidon ». Elle délimite une zone de transition, dite vermillon. Celle-ci se caractérise par sa haute teneur en un liquide semi-fluide qui augmente la transparence du tissu, à tel point qu'on aperçoit la teinte rouge de la couche vasculaire sous-jacente. C'est cette caractéristique qui fait ressortir la couleur des lèvres par rapport au reste de la peau. La zone de vermillon de la lèvre supérieure montre, en son milieu, une protubérance : le tubercule.

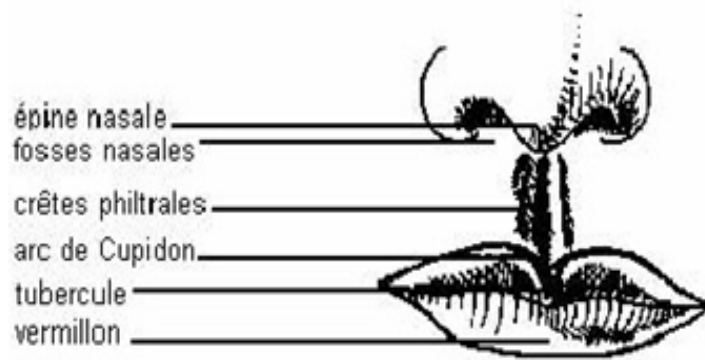


Figure 01.4 – Aspect schématique des lèvres (d'après Zemlin, 1968).

A l'intérieur de la bouche, la muqueuse de teinte rosée rejoint les arcades alvéolo-dentaires. L'espace incurvé, ainsi délimité, forme les gouttières vestibulaires. Dans leurs parties médianes, les gouttières vestibulaires supérieure et inférieure présentent un repli muqueux : le frein de la lèvre. Celui-ci est nettement plus proéminent pour la lèvre supérieure.

1.2.2 Les muscles des lèvres

Les muscles des lèvres font partie des muscles faciaux. Ils ont tous la particularité de présenter une insertion mobile cutanée. C'est cette caractéristique qui rend possible les différentes combinaisons d'expression du visage et la souplesse des mouvements en production de la parole. Le muscle essentiel des lèvres est l'orbiculaire des lèvres qui opère comme un sphincter annulaire. Autour de celui-ci, rayonnent les autres muscles de la face dont les fibres s'imbriquent directement avec celles de l'orbiculaire.

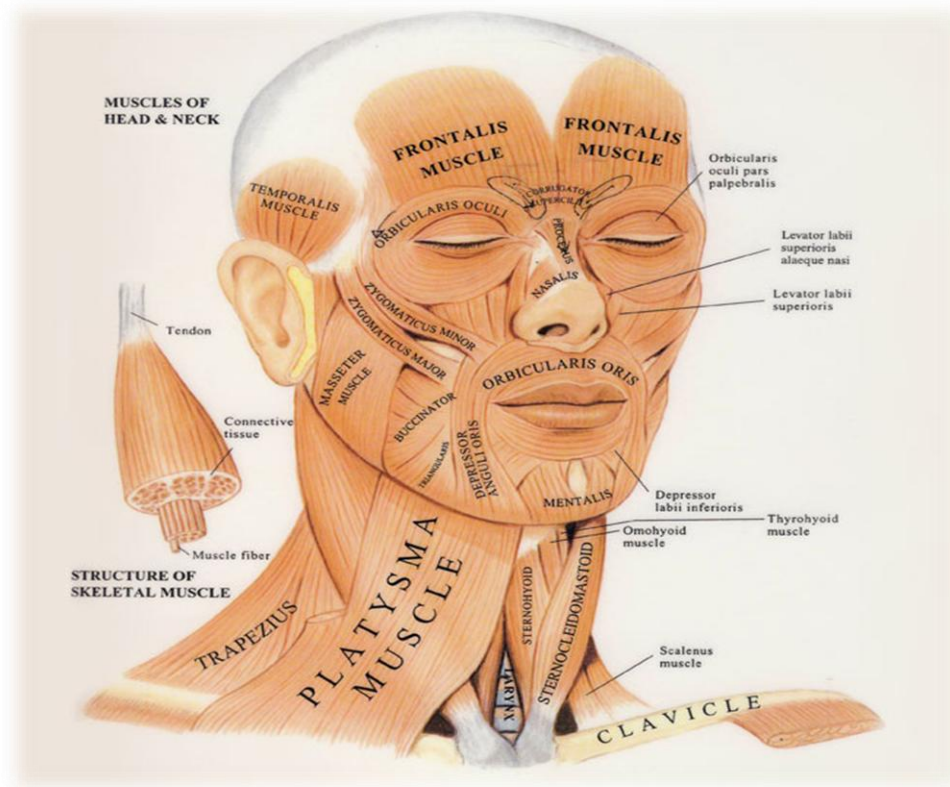


Figure 01.5 – Les muscles de la face (d'après Bouchet et Cuilleret 1972).

Les classifications courantes dénombrent douze muscles pour les lèvres (Zemlin 1968 ; Hardcastle 1976) :

- l'orbiculaire des lèvres (orbicularis oris),
- le canin (levator anguli oris),
- le buccinateur (buccinator),
- les muscles de la houppe du menton (mentalis),
- la carré du menton (quadratus labii inferioris, ou depressor labii inferioris),
- le releveur superficiel de l'aile du nez et de la lèvre (levator labii superioris alaeque nasi),
- le releveur profond (levator labii superioris),
- le petit zygomatique (zygomaticus minor),
- le petit zygomatique (zygomaticus minor),
- le grand zygomatique (zygomaticus major),
- le risorius,
- le triangulaire des lèvres (depressor anguli oris).
- le peaucier du cou (muscle platysma).

1.2.3 Classification fonctionnelle des muscles labiaux

En complément d'études anatomiques, des mesures par électromyographie ont permis de dresser une classification des muscles labiaux suivant les mouvements qu'ils génèrent. Cette classification suit celle de (Hardcastle 1976), reprise dans (Abry 1980). Elle présente les tendances générales observées chez plusieurs sujets.

Muscles assurant l'occlusion labiale

Par contraction l'orbiculaire accole les lèvres supérieures et inférieures en abaissant la lèvre supérieure et en tirant la lèvre inférieure vers le haut. Le mouvement de la lèvre inférieure est fortement dépendant de la mâchoire. Le canin et le triangulaire peuvent aussi intervenir pour fermer les lèvres.

Muscles assurant la protrusion des lèvres

La protrusion correspond à un mouvement poussant les lèvres vers l'avant, s'accompagnant d'un rapprochement des lèvres et des commissures. C'est aussi une des fonctions principales de l'orbiculaire. La houppe du menton contribue à faire basculer la lèvre inférieure.

Muscles assurant l'arrondissement des lèvres

L'arrondissement correspond à une forme de lèvres obtenue en rapprochant les commissures. Ce geste s'oppose à l'étirement. Bien que l'arrondissement s'obtienne par une contraction de l'orbiculaire, ce geste ne s'accompagne pas forcément d'une protrusion. Des muscles comme le buccinateur ou le risorius peuvent limiter l'action de l'orbiculaire.

Muscles élévateurs de la lèvre supérieure

Comme leur nom l'indique, les releveurs supérieurs et profonds de la lèvre sont attachés à cette fonction. Du fait de leur insertion, c'est essentiellement la partie centrale de la lèvre supérieure qui est relevée.

Muscles abaisseurs de la lèvre inférieure

La lèvre inférieure est tirée vers le bas par le carré du menton. Ce muscle peut être aidé par la mâchoire. De même, le triangulaire peut aussi intervenir pour abaisser la lèvre inférieure.