Phonetic sciences

A short introduction

Jacqueline Vaissière

Textbooks in Language Sciences

language
science
press

Textbooks in Language Sciences

In this series:

1. Müller, Stefan. Grammatical theory: From transformational grammar to constraint-based approaches.
2. Schäfer, Roland. Einführung in die grammatische Beschreibung des Deutschen.
3. Freitas, Maria João & Ana Lúcia Santos (eds.). Aquisição de língua materna e não materna: Questões gerais e dados do português.
4. Roussarie, Laurent. Sémantique formelle: Introduction à la grammaire de Montague.
5. Kroeger, Paul. Analyzing meaning: An introduction to semantics and pragmatics.
6. Ferreira, Marcelo. Curso de semântica formal.
7. Stefanowitsch, Anatol. Corpus linguistics: A guide to the methodology.
8. Müller, Stefan. Chinese fonts for TBLS 8 not loaded! Please set the option tblseight in main.tex for final production.

ISSN: 2364-6209

Phonetic sciences

A short introduction

Jacqueline Vaissière

language
science
press

Vaissière, Jacqueline. 2021. *Phonetic sciences*: *A short introduction* (Textbooks in Language Sciences). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/000

ISBN: (Digital)
(Hardcover)

ISSN: 2364-6209
DOI:
ID not assigned!

Cover and concept of design: Ulrike Harbort
Fonts: Libertinus, Arimo, DejaVu Sans Mono
Typesetting software: XƎLATEX

Language Science Press
xHain
Grünberger Str. 16
10243 Berlin, Germany
langsci-press.org

Storage and cataloguing done by FU Berlin

Freie Universität Berlin

Contents

Translator's foreword

The present book is a translation of the fourth edition of Dr. Jacqueline Vaissière's *La Phonétique* published in the Que-Sais-Je Collection by the Presses Universitaires de France (1st ed. 2006; 2nd ed. 2011; and 3rd ed. 2015). The author is a distinguished professor emeritus of linguistics at the University Sorbonne Nouvelle (Paris 3), who continues to teach courses in spectrogram reading and acoustic phonetics and provide training to graduate students in several Parisian universities. Her works have had an important impact on research in all aspects of phonetics (segmental and prosodic) both in terms of empirical approach, speech technologies and theoretical outlook.

The fourth edition of *La Phonétique* has undergone some slight changes in order to add more information or to increase clarity concerning certain points, but the basic structure of the book has remained unchanged. It includes the same topic headings in the same spirit of clarity, avoiding ambiguous issues or technical terms.

In this translation, the bibliography is updated with additional references, and an index is included. The hand-drawn spectrograms in the chapter on prosody have been replaced by narrow-band spectrograms constructed with the use of WinPitch software. The choice of English equivalents for certain French technical terms has been made with the author's permission. The author has also approved the quality of the spectrograms. No notes have been added in the translation.

Dr. Vaissière writes that her book is intended as "both an introduction to the diversity of the phonetic sciences and a synthesis of the results of the research of the last decades," with emphasis on the "renewal of research problems that have accompanied the development of new techniques including computer tools and exploratory techniques." Her short but insightful book provides the reader with a clear overview of the field of phonetics. The nine chapters of the book are carefully structured to present a well-chosen series of topics in an appropriate order that provides the information essential for a basic understanding of the field of phonetics. On the whole, this is a highly recommended textbook for courses in general phonetics (especially at the advanced level) and is also an excellent complement to other textbooks in the field. Dr. Vaissière's *Phonetics* provides a rich hunting ground for graduate students and specialists alike who take a theoretical and empirical interest in phonetics and wish to explore the implications of the issues raised here using experimental techniques, recent methods, and new research tools. This book provides an entryway into an empirically-based detailed phonetic description of speech data in terms of their acoustic properties and perceptual attributes.

I would like to thank Marie Bolton (Associate Professor at UCA), Susan Triebert (my former graduate student, who is now a professional translator), and Sam Cannarozzi (professional story teller) for patiently giving their time to read the translation and offering important refinements. I am also grateful to two of my former graduate students: Ania Akbal, who provided graphic and Photoshop assistance with the figures and Sandra Chassan, who generously lent her fine voice to create the spectrograms.

Readers can visit the PSN website at …… to download the sound files corresponding to the spectrograms presented in this book.

Kambiz ELHAMI

Associate Professor, University of Clermont Auvergne, France

Author's foreword

The subject matter of phonetics is the scientific study of speech sounds. It deals with all the sound phenomena related to the oral expression of human language. The beginning of articulatory phonetics and orthoepy dates to Panini's description of Sanskrit in the 6th century BC.

The 19th century marks the beginning of historical phonetics, with the discovery of phonetic correspondences that bear witness to the interrelationship between languages: the kinship between the languages of Oceania and also between the languages of the vast Indo-European family has thus been recognized. The comparison of related languages leads to the reconstruction of ancient language states, which, as research becomes available, enriches it in a dialogue with historical and paleological data.

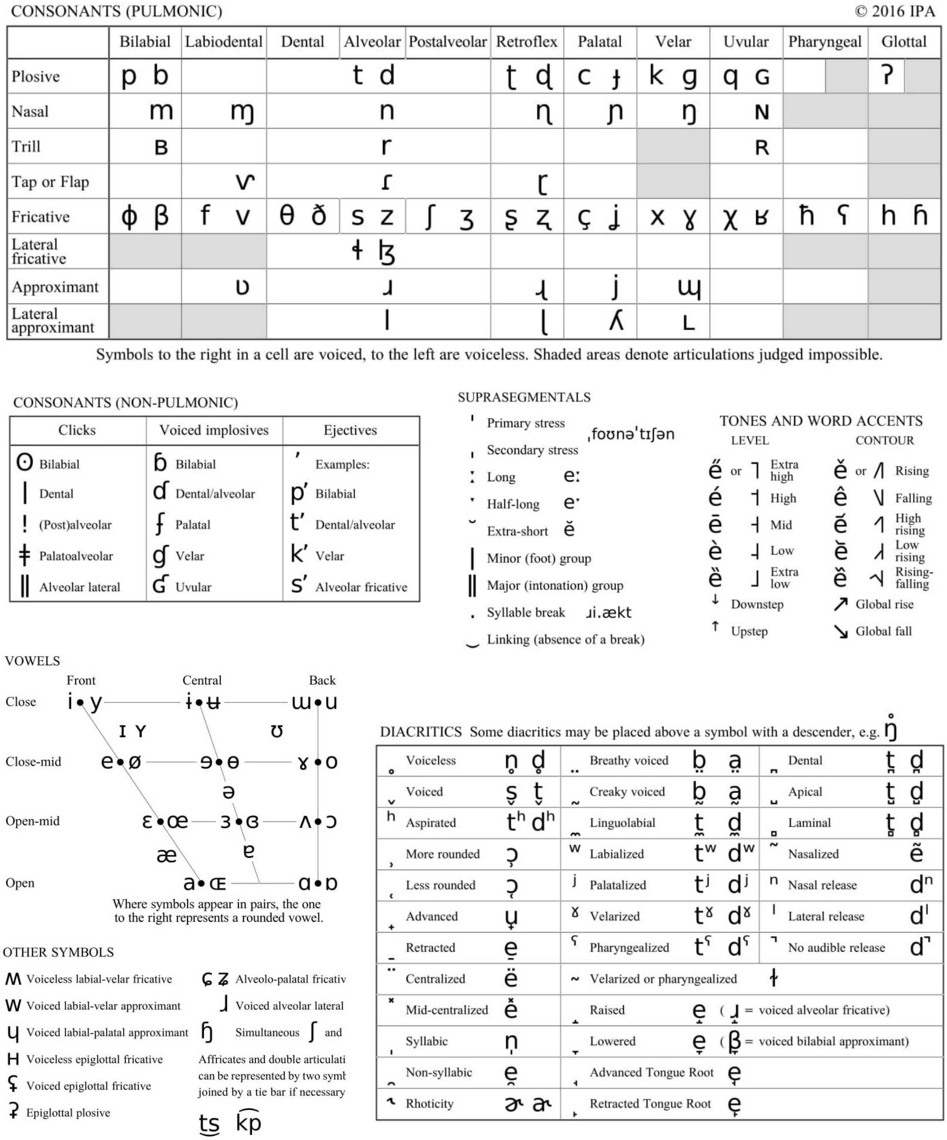
At the end of the 19th century, Pierre-Jean Rousselot tried to explain the mechanisms of phonetic changes through laboratory experiments and founded *experimental phonetics*, a discipline that grew considerably during the second half of the 20th century and has since become a multidisciplinary science with a high usage of specialized instruments.

The first *Congress of Phonetic Sciences* was held in 1932, in Amsterdam. This congress continues to periodically bring together linguists (including phoneticians, phonologists and dialectologists), psycholinguists and experimental psychologists, engineers specializing in spoken communication and automatic speech processing, ENT physicians, voice therapists, speech therapists, neurophysicians, specialists in first-language acquisition and second-language learning, voice and singing teachers, and communication specialists. The collaboration between disciplines is at the origin of major advances in phonetics. Phonetic sciences now concern all scientists, phoneticians or non-phoneticians, whose area of interest is spoken communication: its acquisition, nature and (dys)function.

This book provides both an introduction to the diversity of phonetic sciences and a synthesis of the results of the research of the last decades. Among other limitations due to format, historical phonetics is not presented in detail. The emphasis is on the renewal of the research problems that have accompanied the development of new techniques including computer tools and new exploratory techniques.

Untrained readers are recommended to begin with Chapters 4 and 5.

Table 1: The International Phonetic Alphabet (revised to 2016)



CONSONANTS (PULMONIC) © 2016 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	ʼ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ʄ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɠ Velar	kʼ Velar
ǁ Alveolar lateral	ʛ Uvular	sʼ Alveolar fricative

SUPRASEGMENTALS

ˈ	Primary stress	ˌfoʊnəˈtɪʃən
ˌ	Secondary stress	
ː	Long	eː
ˑ	Half-long	eˑ
̆	Extra-short	ĕ
|	Minor (foot) group	
‖	Major (intonation) group	
.	Syllable break	ɹi.ækt
‿	Linking (absence of a break)	

TONES AND WORD ACCENTS

LEVEL		CONTOUR	
e̋ or ˥	Extra high	ě or ˩˥	Rising
é ˦	High	ê ˥˩	Falling
ē ˧	Mid	e᷄ ˦˥	High rising
è ˨	Low	e᷅ ˩˨	Low rising
ȅ ˩	Extra low	e᷈ ˧˦˧	Rising-falling
ꜜ	Downstep	↗	Global rise
ꜛ	Upstep	↘	Global fall

VOWELS

	Front		Central		Back
Close	i • y		ɨ • ʉ		ɯ • u
		ɪ ʏ		ʊ	
Close-mid	e • ø		ɘ • ɵ		ɤ • o
			ə		
Open-mid		ɛ • œ	ɜ • ɞ		ʌ • ɔ
		æ	ɐ		
Open		a • ɶ			ɑ • ɒ

Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative

w Voiced labial-velar approximant

ɥ Voiced labial-palatal approximant

ʜ Voiceless epiglottal fricative

ʢ Voiced epiglottal fricative

ʡ Epiglottal plosive

ɕ ʑ Alveolo-palatal fricativ

ɺ Voiced alveolar lateral

ɧ Simultaneous ʃ and

Affricates and double articulati can be represented by two symb joined by a tie bar if necessary

t͜s k͡p

DIACRITICS Some diacritics may be placed above a symbol with a descender, e.g. ŋ̊

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	ʲ	Palatalized	tʲ dʲ	ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	ˠ	Velarized	tˠ dˠ	ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	◌̴	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)		
◌̩	Syllabic	n̩	◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
◌̯	Non-syllabic	e̯	◌̘	Advanced Tongue Root	e̘			
˞	Rhoticity	ɚ a˞	◌̙	Retracted Tongue Root	e̙			

Introduction

Phonetics is the study of the *substance and form* of speech sounds. This deliberately broad definition will be thoroughly explained as we move forward in our presentation, comparing this discipline to its related fields.

The ability to acquire a language is unique to humans, and as such, a hallmark of the human species. Animals, even the most primitive, have systems of olfactory, visual or auditory clues that allow the exchange of information with their fellow creatures for the survival of their species. The inventory of these clues is limited in number and this is not strictly speaking a matter of language. In this regard, Emile Benveniste separates *animal communication* from *human language*. Human beings have a faculty that enables them to make an *unlimited* number of messages, contrary to animals, whose verbal exchanges would be *limited* in number, Benveniste (1953).

The ability to acquire a language is manifested in an extraordinary diversity of spoken languages, ranging from 3000 to about 6500, depending on the criteria used to count them. Speech as the main way of communication may also use, in a controlled way, other *noises* made by the same organs as for speech (sighs, laughter, coughing, and marginally onomatopoeia) or the information provided by the rest of the body, where gestures and facial expressions make a postural mimo-gestural system. Deaf children often spontaneously develop a communication system among themselves with visual signs. Humans truly are creatures prone to communicating.

Spoken language provides a wide variety of information that is not present in written language. Depending on how the speaker pronounces a verbal message, he communicates extra information in various forms as compared to the written message, in a way that is only partly under his control. He externalizes his feelings about what he says (conviction, doubt) and about his interlocutor (disdain, respect), emotions or attitudes (joy, distress, fear). He expresses the need for a particular reaction from his interlocutor (response to a question, an order). He also reveals his social, regional and cultural identity. *Phonostylistics* (a term used by Ivan Fónagy and Pierre Léon) deals with these dimensions, present in every act of communication and particularly prominent in the aesthetic use of voice (song, poetry, performing arts).

Language can be described in terms of *double articulation* (a term originally used by André Martinet). Each message is composed of a sequence of *signs*, and each sign corresponds to a succession of elementary sounds. A sign (typically a word or a morpheme) has two sides, a *signifier* (sound image) and a *signified* (meaning). The correspondence between the signifier and the signified is *arbitrary* and *conventional*: arbitrary, because the notion of *tree* is designated by a different sound image depending on the language, *arbre* [aʁbʁ] in French, *tree* [triː] in English, *Baum* [ˈbaʊ̯m] in German. Conventional, because speech is "a social product of the language faculty and a set of necessary conventions adopted by society to allow the use of this faculty in individuals" (Ferdinand de Saussure). Each signifier is itself made up of a sequence of meaningless elementary sounds, called *phonemes*. The combinations of the three phonemes /p/, /t/ and /a/ will make at least five meaningful units in French (*pas* "not", *ta* "your", *patte* "paw", *tape* "present indicative form of the infinitive *taper* 'to type', used with the first or second person subject", *apte* "able": /pa/, /ta/, /pat/, /tap/, /apt/). The same sequence of the two phonemes /sɑ̃/ correspond to several words (*sang, sans, sent, cent, s'en*, meaning "blood", "without", "smells", "hundred", "in", respectively) , Martinet (1949). Each language distinguishes about tens of thousands of words made from an average of 30 phonemes.

The phoneme is the smallest functional unit of the phonological system. The function of phonemes in a language is to auditorily distinguish one word from another. "If two sounds occur in exactly the same phonetic environment and cannot replace each other without changing the meaning of words, or without rendering the word unrecognizable, then they are the realizations of two phonemes" (Nikolaj Sergeyevich Troubetzkoy, Troubetzkoy (1949). In French, /l/ and /ʁ/ are two different phonemes because replacing /ʁ/ for /l/ results in two different words, as in *père* (/pɛʁ/) "father" and *pelle* "shovel" (/pɛl/) or *rang* (/ʁɑ̃/) "rank" and *lent* (/lɑ̃/) "slow". However, the Parisian uvular r, pronounced [ʁ] and the rolled r [r] (called *the Burgundy* [r]) are two regional variants of the phoneme /ʁ/. Two words that differ only in one phoneme constitute a minimal pair: *lent* (/lɑ̃/) "slow" et *rang* (/ʁɑ̃/) "rank" make a minimal pair establishing the phonemic status of /l/ and /ʁ/ in French.

Phonemes include vowels and consonants, as well as semi-consonants and semi-vowels. The number and identity of phonemes vary from language to language. The majority of languages have 25 to 30 phonemes. In the extreme, the Pirahã (the language of the Amazon) has only 10 phonemes, and the !Xũ (in South Africa) more than 100; French has 27-33 phonemes depending on the regions and generations. French has 16 consonants and this number does not vary from a regional variety to another: / p t k b s g f s ʃ v z ʒ m n l ʁ/ (see Table 1 for the list of French phonemes). The number of vowels, however, is variable. Older French speakers, for example, in the region north of the Loire River, have two phonemes of the type /a/, the front /a/ and the back /ɑ/, a distinction that exists only in a few words. For example, they pronounce differently *patte* /pat/ “paw” and *pâte* /pɑt/ “dough”, or Anne /an/ ”Ann” and *âne* /ɑn/ “donkey”. Southern French and the French of French-speaking parts of Belgium distinguish between *brun* /bʁœ̃/ “brown”and *brin* /bʁɛ̃/ “strand”, a distinction that is lost in Parisian French. The words *Baule* /bol/ – a city in France – and *bol* /bɔl/ “bowl”, and *fée* /fe/ “fairy” and *fait* /fɛ/ “fact” are homophones in Normandy. Standard French no longer distinguishes between long and short vowels. In Belgium and South Lorraine (Nancy, Saint-Dié), however, the length contrast still allows the distinction between *mot* [mɔ] “word” and *maux* [moː] “aches, pains” (opposition coupled with a slight difference in timber, as is common for the vowel length oppositions) but the trend for a change to a [o] sound in open syllables in word-final position and [mɔ] in close syllables in word-final position is spreading rapidly. [o] and [ɔ] are then two phonetically distinct variants (allophones) of the single phoneme /o/. As the Media commentators are required to have a pronunciation of a *neutral type*, corresponding to what is supposed to be the standard variety of the French language, we observe a tendency towards standardization.

The phonetic realizations of a phoneme vary enormously depending on different factors.

Firstly, the variations are due to individual anatomical characteristics (length and shape of this vocal tract, physiological characteristics of his vocal folds) and give information about the speaker as to his age, sex, physiological state (hoarse voice, smoking voice) and emotional state (happy voice, sad voice).

Secondly, the phonetic context surrounding the phoneme influences the movements of the tongue and lips, the soft palate and the vocal folds that produce it. A phoneme is not realized in an identical manner in all environments. Compare the position of your lips during the realization of the consonant /t/ in *toute* /tut/ ”all” [t^{w}ut^w] (/t/ is said to be contextually labialized or rounded) and *tête* /tɛt/ ’head”. The lips are more protruded and rounded during the production of the initial and final consonants in *toute* /tut/ than in *tête* /tɛt/. Compare also the position of your tongue during the realization of the consonant /t/ in *toute* /tut/ ”all” and *tête* /tɛt/ ’head”. The tongue are more back during the production of the initial and final consonants in *toute* /tut/ (/t/ is said to be contextually retracted) than in *tête* /tɛt/. Lip rounding and tongue retraction result in a cavity in front of the /t/ constriction shorter in the case of the /t/’s in /tut/ than in the case of the /t/’s in /tɛt/. The explosion noise at the time of the separation of the tongue and the teeth (that is to say at the moment of release of the consonants) has

Table 1: — The list of French phonemes

CONSONANTS							
/pɑ̃/	pan	/bɑ̃/	banc	/fɑ̃/	faon	/vɑ̃/	vent
/tɑ̃/	temps	/dɑ̃/	dent	/sɑ̃/	sang	/zɑ̃/	zan
/kɑ̃/	camp	/gɑ̃/	gant	/ʃɑ̃/	chant	/ʒɑ̃/	gens
/mɑ̃//	ment	/lɑ̃//	lent				
/nɑ̃//	nan	/ʁɑ̃//	rang				

VOWELS					
in closed syllables					
/pil/	pile	/pyl/	pull	/pul/	poule
		/ʒøn/	jeûne	/pol/	pôle
/pɛl/	pelle	/ʒœn/	jeune	/pɔl/	Paul
/pat/	patte			/pɑt/	pâte
/pɑ̃t/	pente	/pɑ̃t/	ponte		
/dɛ̃d/	dinde				
in open syllables					
/li/	lit	/ly/	lu	/lu/	loup
/le/	les	/lø/	leu	/lo/	l’eau
/lɛ/	laid	/bʁœbi/	brebis		
/la/	la	(/lə/)	(le)		
/pɑ̃/	paon	/pɔ̃/	pond		
/bʁɛ̃	brin	/bʁœ̃/	brun		
Semi-vowels (also called semi-consonants)					
/fij/	fille	/lɥi/	lui		
/jɔt/	yacht	/lwi/	Louis		

Check textstylemwheadline for vowels

therefore a lower frequency range for the /t/'s in /tut/ than for the /t/'s in /tɛt/ and is consequently perceived as more grave (the longer the cavity, the more grave the resonances due to that cavity). Furthermore, the tongue is in a more fronted and less lowered during the realization of the phoneme /u/ in *toute* /tut/ "all" than during that of *Ruhr* /ʁuʁ/ "Ruhr" (/ʁ/ is uvular, i.e. articulated with the back of the tongue near the uvula). When extracted from its phonetic context, the timber of the sound contextually fronted [u] in *toute* corresponding to the realisation of the phoneme /u/ resembles that of the front vowel /y/, because the tongue is in a fronted position during the production of the vowel, whereas the contextually retracted and opened sound [u] in *roure* is similar to that of /o/, because the tongue is in a backed and more open tongue position during the vowel. The influence of a phoneme can be felt within the whole word. Take, for example, the influence of the vowel of the final syllable in the words *phonologie* /fonoloʒi/ "phonology" and *phonologue* /fonolɔg/ "phonologist." A native speaker will tend to pronounce the word *phonologie* phonetically as [fonoloʒi], using the high-mid back allophone [o] of the phoneme /o/ three times because the final vowel /i/ in this word is a high vowel and it influences all the preceding vowels in the word, and *phonologue* [fɔnɔlɔg] with the mid back allophone [ɔ] of the phoneme /o/ because of the influence of the final mid back vowel [ɔ] on the other vowels. The word-final high vowel /i/ tend to close all the preceding vowels, and the word-final mid close vowel /ɔ/ to close all the preceding vowels.

Thirdly, the more one speaks rapidly and/or in a relaxed way, he spends less effort to realize the contrast between successive phonemes. The articulatory (and auditory) distance between vowels and consonants in sequence decreases and the coarticulation between successive sounds in sequence increases: the successive sounds are more coarticulated. In relaxed speech, the low (open) vowels with pronunciation requiring a very open vocal tract (called *low vowels*) like the vowels /a/ (as in *patte*), /ɑ/ (as in *pâte* "dough") or /ɑ̃/ (as in *pente* "slope") tend to close, i.e. to be pronounced with a less open vocal tract, and on the contrary, high (close) vowels such as/i/ (as in *pile*) "battery", /y/ (as in *pull*) "pull" and /u/ (as in *poule*) "hen" tend to open, i.e. to be realized with a more open vocal tract; extreme tongue position for the realization of vowels such as /i/, /u/ and /a/ are avoided (oui [wi] "yes" > ouais > [wɛ] (opening of the close vowel /i/ due to a reduced effort to utter the word) and [ɥɛ]) (fronting of /u/ due to the influence of following front vowel /i/). Each phoneme may be influenced by adjacent or nearby segments or/and influence adjacent or nearby segments. Stop consonants (the vocal tract for canonical stops is completely obstructed, as in the realization of /p, t, k, b, d, g/) are often no longer realized with a complete constriction of the vocal tract. In rapid and/or relaxed speech, some phonemes almost disappear or disappear completely: *je ne sais pas* “I don’t know” */ʒœ nœ sɛ pa/ [ʒənəsɛpa]* > *je n’sais pas [ʒnəsɛpa]* > *j’sais pas [ʒsɛpa]* > *ch’sais pas [ʃsɛpa]* > *ch’pas [ʃpa]*; *maint’nant* [mɛtnɑ] > *maind’nant* [mɛdnɑ̃] > *mainnant* [mɛnnɑ̃] > *mai-nant* [mɛnɑ̃] common in children’s speech. These phenomena of reduction are not unique to French, as shown in the experimental work of Klaus Kohler on German, Kohler (1990), for example. These coarticulatory phenomena are often phonologized (transformed) into sound changes over time. Common words are often pronounced more quickly and in a more relaxed way than rare words, and it is through them that phonetic changes begin, which then spread gradually to the less frequent words (the principle of lexical diffusion).

Fourthly, the articulatory gesture and degree of resistance to coarticlation depend on the status of the syllable in terms of prominence (stressed, unstressed or reduced), the syllable position in the word (including initial, final, and intermediate positions), the phoneme position in the syllable (onset or coda), and the word position in the utterance. The phoneme in a strong position in the word is *dominant*, i.e. more prototypical, better articulated, longer, more intense, and unlikely to undergo a sound change or disappear; moreover, it will impose some of its features on the surrounding phonemes. Otherwise, it is *dominated* and under the influence of its surrounding phonemes (for example, it is weakened like the /t/ in city /ˈsɪ**t**i/ > [sɪ**t**i] > [sɪ**r**i]) and is sometimes on the verge of disappearing (like the /t/ in pan**ts** /pæn**ts**/ > [pæns]). The first consonant in the word and in the stressed syllable usually has a strong realization, that is to say it is more prototypical than the other consonants in the word. In English, the stops /p/, /t/ and /k/ are aspirated when they are in word initial-position (pin [p^{h}ɪn], ˈcontract [ˈk^{h}ɒntræt] and conˈtract [k^{h}ənˈtræt]) and when occurring before the stressed vowel of the word, but not elsewhere (toˈma**to** [t^{ə}ˈmeɪ**to**ʊ]); the symbol ˈ indicates the position of the stressed syllable. Consonants in syllable-initial position are better pronounced and better perceived than those in the coda and are more likely to remain. In an utterance, the initial syllables (and words) are generally better articulated than the final syllables. A consonant in a coda position, or a syllable after a stressed vowel, is in a particularly weak position. The part of the statement located after a focus is often pronounced with reduced effort. The augmentation of effort on one element (phoneme, morpheme, word) is done to the expend of the effort spend on the other elements, in particular on the following ones.

Fifthly, the same speaker adapts his way of speaking to the situation of communication, in terms of register (from the most formal to the most familiar), style and expressed attitude (pouting, irony), making use of any means. In this regard, nasalization conveys a mark of respect in some languages and disgust in others. Variants can also be sociolinguistic or sociocultural: fronting and lengthening of the vowel /a/ into [ɑː] in a word like *mariage* immediately indicates the speaker's social identity (the so-called accent of Marie-Chantal of the sixteenth Parisian district; see the work of Philippe Boula de Mareüil for French examples).

Sixthly, the phonemes and their contrasting counterparts are realized differently from one region to another: the Southern *singing* accent is easily recognizable; more subtle differences make it possible to distinguish between the accents of Lyon and Grenoble.

Finally, voluntary changes in the way one articulates the sounds bring shades of meaning to the message (Fónagy (1983): the way a sentence is pronounced, with softness, kindness, coldness or contempt, plays an important role in human interactions. These variations can bring about a sharp change in the overall meaning of the message: a hyper-articulation of the consonant /s/ with lengthening and increased articulation effort in the utterance *elle est sympa!* "She's nice!" means that the person is anything but friendly.

The human baby is predisposed to speak. The fetus absorbs at an early stage the verbal sounds and rhythm of the mother tongue perceived through the amniotic fluid. A baby of a few days reacts to the phonemic oppositions of almost all languages of the world and not just those of his mother tongue, but he already distinguishes the language of his mother from other languages. English adult speakers do not differentiate between a dental stop (articulated with the tip or blade of the tongue making contact with the teeth during production) and a retroflex (the tip of the tongue is directed upwards and towards the back of the mouth) or between voiced aspirated and unaspirated voiced stops in Hindi (a contrast which is absent in English). Babies and future speakers of English or other languages, however, can perceive the difference between these different sounds shortly after birth. After initial babblings, at around the age of 6 months, the baby already mimics the sounds and intonation of the people around him. The lack of such a gift of imitation would be the cause of the inability of monkeys to learn to articulate a wide range of sounds. Recent experiments with techniques of evoked potentials show that the brain of the child, even when sleeping, at around the age of 8 months, reacts differently to sound contrasts whether they are used in his mother tongue (phonemic contrasts) or not. The brain of the French baby reacts to the difference between the sounds corresponding to the two phonemes /i/ (with spread lips) and /y/ (with rounded lips), whereas the brain of an English baby does not (English does not use the lip rounding feature distinctively). Very quickly, the baby is able to store the acoustic forms to which he is frequently exposed, well before he understands them. Very early, the baby has a preference for phoneme sequences frequently occurring in his mother tongue. Around the age of 8 or 10 months, the child gradually becomes indifferent to contrasts between non-pertinent sounds in the spoken language around him, holding only those offered by the mother tongue. Chinese babies of a few days have no trouble distinguishing between [do] and [to], [ga] and [ka] (for which the voicing feature is distinctive), while Chinese adults learning French have great difficulty in hearing the difference between *gâteau* [gato] "cake" et *cadeau* [kado] "gift", because *they have learned not to make a difference* during their acquisition of Chinese, which is a language that does not use the voicing feature distinctively. Similarly, Japanese babies between 8 months and 1 year old gradually lose their sensitivity to the difference between /l/ and /r/, and Japanese adults have some trouble hearing the difference between *lit* [li] "bed" and *riz* [ʁi] "rice", *even after long years of exposure to French!* Linguistic experience profoundly influences the perception of sounds (*filter model of attention* according to Janet Werker, Werker & Curtin (2005) , *psychoacoustic reorganization* around specific prototypes of the language according to Patricia Kuhl, Kuhl (1991). Learning a second language (during adolescence or adulthood) requires a sustained effort not to be influenced by the system of his mother tongue and necessitates learning a new phonemic system, i.e. that of the foreign language. An aspiring phonetician too should go through rigorous training in order to distinguish all types of sounds used contrastively in the languages of the world and noted in the IPA (International Phonetic Alphabet), updated by the International Phonetic Association (see Table 1 on page vi). The inventory of possible phonemic oppositions in languages has not been completed, although the discovery of new types of phonemes is becoming increasingly rare.

The results of neurological research seem to support the longstanding innate theory of the faculty of language advocated by Noam Chomsky in the last century according to which the human baby would be born with the ability to acquire a language with double articulation, which is not shared by animals.

The human vocal tract can produce an indefinite number of different sounds, as evidenced by the feats of beatboxers, whose vocal apparatuses can uncannily mimic the sounds produced by musical instruments. In spite of this ability to produce an unlimited number of sounds, there are a very large number of phonetic similarities among the sounds selected by the world's languages. These similarities are due to the presence of the same phonetic (or substantial) constraints imposed by the general laws of acoustics and aerodynamics, by the same characteristics of the systems of speech production and speech perception, and by the human brain structures that generate the same cognitive skills (such as short-term and long-term memories, and the faculty of learning and generalizing). Inspired by the discrimination performance of sounds by newborns, researchers have advanced the idea that humans are equipped with a limited number of *auditory detectors of properties or features*, pre-wired for human speech (an idea developed by Ken Stevens), of which only certain examples are selected to realize oppositions between words within a language.

Transmitting information via an acoustic signal has great advantages over other media. Speech allows interlocutors to be relatively distant. It remains usable in a noisy environment. To use speech *frees* the sight and hands, which can then perform other tasks. It is also a quick mode of communication: a speaker in a hurry can produce more than 30 phonemes per second, more than 200 words on average per minute, and his listener can grasp the message in real time. If the sequences of non-linguistic sounds were presented at the same rate, the listener would perceive only noise. As we will see later, speech sounds are not treated by perception mechanisms in the same way as the sounds of nature.

1 Phonetics and phonology

Phonetics and phonology (the latter also called *functional phonetics*) are two branches of linguistics that study the sound side of language. The sharing of objectives between phonetics and phonology has evolved in successive steps for over a century.

At the beginning of the last century, Ferdinand de Saussure, Cours de Linguistique Générale, De Saussure (1957) characterized language as a system, the elements of which are defined by their relationships to each other and stressed the independence of the study of an abstract linguistic system (the language or form, i.e. the system) and its concrete phonetic realization (i.e. speech or substance). Each element (a phoneme) is defined as a unordered bundle of binary, formal distinctive (contrastive, phonological, abstract) features. In a further step, the representatives of the Prague Circle (including Roman Jakobson and Nikolaj Sergeyevich Troubetzkoy) advised clearly separating the study of sounds (the substance), i.e. the subject of *phonetics*, from the study of the sound system (the form, the organisation), i.e. the subject of *phonology*. Troubetzkoy defined phonetics as *the science of the physical side of the sounds of human language* ,Troubetzkoy (1949). Phonology would only be concerned with phonemic oppositions, namely in the system of oppositions that a language uses. The relation between a phoneme and its realisation is considered as arbitrary. This distinct separation between phonetics and phonology had a positive impact on their separate developments. Phonetics benefited from this division to get close to engineering sciences and life sciences. At the same time, attention focused on the analysis of language *systems* allowed phonology to make undeniable progress.

Some of the most striking works are nevertheless the result of collaboration between linguists and engineers. The book *Preliminaries to Speech Analysis* (Jakobson et al. (1952) co-authored by Jakobson (the great Russian linguist belonging to the Prague Circle) and Gunnar Fant (the famous Swedish telecommunications specialist), marks a turning point in the history of the relation between phonology and phonetics: the notion of distinctive features, previously regarded as a formal notion in phonological analysis, is described in terms of physical features based on the acoustico-perceptual properties of speech sounds, or in other words in terms of the substance (i.e. the physical material) of speech sounds. It should be noted that for Trubetzkoy as well, the distinctive features of speech sounds were derived from their substance, as they were described in articulatory terms. The notion of phonetic constraints was then applied to the description of phonological rules and coarticulation. The idea of the arbitrariness of the relation between the physical nature of sounds and phonological systems is completely abandoned: phonological form and phonetic substance are mutually dependent. The later convergence of phonetics and phonology has been a great factor of progress.

The phoneme inventories in a language may be more or less complex, Ladefoged & Maddieson (1998). Some regularities emerge in the phoneme inventories of languages such as the choice frequency of /i/, /a/, /u/ in phonological systems with three vowels, and that of /i/, /e/, /a/, /o/, /u/ in systems with five vowels (the most numerous cases, including 22% of the languages of the UPSID[1] database). Note that the exact phonetic realization of these few phonemes depends on the language, Disner (1983). /u/ ([u]) is realized as strongly rounded in French, and unrounded in Japanese ([ɯ], sometimes designed as /ɯ/). The rounding of the lips during English /u/ is not strong enough for English /u/ to sound like French /u/.

Phonetics and phonology are both concerned with the definition of all the constituent features of phonemes, concrete or abstract. The common basic scientific questions addressed by the phoneticians and the phonologists ‘are the following: How much is the choice of phonemes in language inventories arbitrary? How do we explain the general inter-language tendency to choose phonemes with close phonetic realizations? Does the choice of a given phoneme determine other phoneme choices? Why and how do phoneme inventories change over time? None of these questions have received a definitive answer. To answer these questions, phoneticians and phonologists are interested primarily in the constraints, in the broad sense, which govern the choice of systems of sound oppositions in world languages and their evolution over time within the same language. For explaining the choice of phonemes in language inventories, theories put more emphasis either on *substance based constraints* or on *formal explanations* (each phoneme correspond to a bundle of binary distinctive features).

According to Jakobson, the distinctive features should have definable phonetic correlates. Jakobson gives priority to the constraints based on the phonetic substance of speech sounds for the choice of phonemes: that is to say, the oppositions between phonemes in a language, for Jakobson, are based on their *acoustic* correlates and *ease of perception* by the listener. For Jakobson, production constraints

[1] The UCLA Phonological Segment Inventory Database of the University of California lists 920 different speech sounds, over 650 consonants and over 260 vowels in 451 languages, Maddieson (1981)

(i.e. *articulatory* constraints) are a secondary concern. Jakobson proposes a definitive list of a dozen *universal* distinctive features of which phonemes are composed (such as *vocalic/non-vocalic, consonantal/non-consonantal, compact/diffuse, tense/non-tense*, etc.). Each language would choose between these pre-existing, innate, features to realize oppositions between words. For example, French uses nasality and labiality to constrast its vowels, German length and Chinese tones.

In their search for formal explanations of alternations for example between [œ] and [ø]) observed in the words *peur* [pœr] "fear" and *peureux* [pørø] "fearful", *beurre* [bœr] "butter" and *beurré* [børe] "buttered", Noam Chomsky and Morris Halle, Chomsky & Halle (1968) , unlike Jakobson, give less importance to the substance-based definition of distinctive features: both authors highlight their formal, abstract, definition, a perspective which is still adopted in some current research in phonology. Distinctive features are thus vaguely defined by Chomsky and Halle in an essentially articulatory way (such as the features ”High” and ”Back”), without providing details about their acoustico-perceptual characteristics or even production constraints.

Among the constraints based on the substance (called phonetic constraints) *anatomical* constraints are addressed: for example, the tip of the tongue allows an constriction of greater precision in change in place of constriction and shape (convex or concave) than the root of the tongue. The consonants selected by the systems of the world's languages are essentially realized with a constriction in the front section of the vocal tract, and the very mobile tip of the tongue, which is particularly put to use for the production of consonants. Some regularities emerge in. Over 99% of languages have the front phoneme /t/. Few languages have pharyngeal consonants, in the production of which the root of the tongue comes into play in a small number of oppositions, Ladefoged & Maddieson (1998).

In the 1970s, the discussion on phonetic constraints was significantly brought back to *perceptual and acoustic planes* (and not the articulatory plane), and two main ideas were advanced.

- 1) certain phonemes be selected because of their *intrinsic properties*. Phonemes like /i/, /a/ and /u/) would be selected according to the stability of their acoustic properties, with the idea that producing the timber of these three vowels would not require the highest degree of precision in articulatory gestures (which would not be efficient). According to Ken Stevens's *quantal theory*,Stevens (1989), which will be seen later, /i/, /a/ and /u/ would be preferred because they are acoustically ”quantal” vowels.

- 2) phonemes are not only chosen for their intrinsic properties but also for their capacity to perceptually distinguish those phonemes that are acoustically close to them. The entire phonological system of vowels or consonants would have an influence on the individual choice of vowels and consonants, especially when their number in the language is high, and the distinctive sounds would tend to position themselves in the perceptuo-acoustic space in a manner to maximize their mutual perceptual contrast (Bjorn Lindblom's *dispersion theory*, Lindblom (1986), so that the phonemes are easily differentiated perceptually. For example, voiceless nasal consonants are often perceived as voiceless fricatives; hence there are very few cases of phonological opposition between voiceless and voiced nasal consonants and few cases of voiceless nasals in languages (Bhaskararao & Ladefoged 1991). [+nasal][-voiced] consonants are somewhat difficult to distinguish perceptually from [+fricative] [-voiced] consonants.

Both theories are combined in the *dispersion-focalization theory* , put forward by the GIPSA-LAB of Grenoble (Schwartz et al. 1997). According to dispersion-focalization theory, the phonological system of a language is the result of a particular balancing of the language between the dispersion of phonemes in the acoustico-perceptual space (i.e. dispersion) and the acoustico-articulatory stability of each phoneme (i.e. focalization).

- 3) certain *combinations of features* defining a phoneme are avoided in many languages because they are difficult for the speaker to produce and for the listener to perceive. For example, rounding and protrusion of the lips is very difficult when the vocal tract is wide open, i.e. for very low vowels, hence there are very few cases of phonological opposition between low vowels made with rounded lips and those with spread lips). The simultaneous realization of the abstract features [+round] and [+low] is somewhat difficult to realize on the production side.

- 4) *perceptual constraints* on a phoneme inventory can be also in the form of phoneme selection in a *sequence of phonemes* : a speaker may have difficulty effectively realizing a combination of features for two successive sounds that will therefore be poorly identified by the listener and may be avoided in a language.

  The sequences /ji/, /ɥy/ and /wu/ are avoided because the acoustic distance between the vowels and their approximant counterparts is very small, and therefore difficult to segment. Some sequences of sounds are easier to pronounce than others and this can result in restricting the number of possible syllables: syllables composed of phonemes, whether they are all anterior phonemes like /ti/ or posterior like /ʁu/, are easier to pronounce than /tu/

and /ʁy/ with combinations of anterior and posterior phonemes. Just think of the difficulty of English speakers in distinguishing between *Les russes sont rousses* "Russians are red-haired", and *les rousses sont russes*! "The redheads are Russian."

In some languages, these constraints have been *phonologized* over time and some sequences of phonemes in a word are not allowed as in the case of *vowel harmony*. In Turkish, words may not contain both front and back vowels (so called long distance vowel harmony, (Krämer 2008).

Perception also plays an important role in this regard. Bilabial fricatives rarely persist: the human ear cannot perceive them well, especially in non-ideal acoustic conditions, which are those of everyday communication with surrounding noises. For example, the bilabial /ɸ/ and /β/ are often replaced by the more audible labio-dental /f/ and /v/). Nasal consonants, such as /m/ and /n/ in coda are difficult to be perceptually distinguished when they are unreleased. The distinction between /n/ and /m/ in coda is not relevant in Standard Chinese and Japanese unlike in English and French. The number of nasal vowels is always equal to or less than that of oral vowels in a language: the acoustic correlates of nasality make it more difficult to distinguish between the different vocalic timbers of nasalized vowels, and this difficulty tends to reduce the number of oral-nasal vowels over time in a language, (Beddor 1993). For example, there are only four nasal vowels left in French, and the distinction between /ɛ̃/ and /œ̃/ tends to disappear in Modern French (but not in South France, Belgium and Canada). The perceptual contrast between two phonemes in a non-mother language may be difficult to perceive. Some languages have subtle phonemic oppositions, for example, the opposition between dental and alveolar stops (Jongman et al. 1985). and the clicks of Zulu (distinctively dental, alveolar or lateral) is very difficult to distinguish by an untrained ear (Best et al. 1988). But language listeners who differentiate between dental and alveolar stop consonants, or Zulu natives, have no difficulty in perceiving and producing phonemic oppositions contained in their respective inventories, and they do so from a very early age.

- 5) *aerodynamic constraints* play a large in the selection of the phonemes contrasting by the voicing feature. Among the constrictives (including stops and fricatives), the voiced examples (such as /b, d, g, v, z, ʒ/) are less frequent than the voiceless examples (/p, t, k, f, s, ʃ/) for *aerodynamic* reasons: voicing is a disadvantage when the intraoral pressure is high. The intraoral pressure rises in the case of closing (such as for stops) or narrowing (such as for fricatives) of the vocal tract. Try to maintain voicing as long as possible during the production of the consonants / b / and / g /. The maintenance of the voicing is facilitated in the case of / b / as compared to / g /: the swelling of the cheeks makes it possible to expand the volume of the cavity in front of the constriction and to maintain the passage of an air flow through the vocal folds, (Ohala 1983). Voicing is particularly disadvantaged in the case of a posterior constriction, where the cavity behind the narrowed area can hardly be extended: as noted previously, /g/ is rare in languages but it can be maintained for phonological reasons, as symmetry of the consonant system, i.e. /p, t, k/ opposing /b, d, g/.

- 6) *visual constraints* also play an important role and provide some explanations about certain observed facts. The first consonants acquired by seeing babies, but not by blind babies, are the bilabials (/p, b, m/ followed by /n, t, k, g/), and this is evidence of the importance of viewing the speaker's face; often, a baby with his eyesight intact fixes his eyes on his mother's lips when she addresses him, (Kuhl & Meltzoff 1982). Speech in a noisy environment is better recognized if the listener sees the speaker's face, which is further evidence of the importance of visual cues (Stein et al. 2009). Visual information on place of articulation easily influences phonetic perception, (McGurk & MacDonald 1976).

- 7) *cognitive constraints* such as ease of learning and memorization, also play an important role: they favor a reduction in the number of distinctive features used in a language and their organization into an efficient and symmetrical system, with maximum use of phonological features chosen, maximising the ratio of sounds over distinctive features by the language to realize sound oppositions between words (Martinet 2020), (Clements 2003). In French, the correlation of voicing features (voiced-voiceless opposition in consonants) allows oppositions in the symmetrical series /p/, /t/, /k/, and /b/, /d/, /g/. For example, /p/ is absent in Arabic, and /g/ in Dutch: voicing is more difficult to maintain for the production of /g/ than for that of /b/ or /d/.

- 8) *external factors (historical, geographical, sociological, language contacts, bilingualism* also play a significant role. In this respect, contact between languages and imitation within a language of a variety considered more prestigious, sex, age, can also be a source of change (see the work of the sociolinguist William Labov, (Labov 1972).

The study of phonetic constraints on inventory systems and on phonetic changes has a traditional background in phonetics (Pierre-Jean Rousselot as a pioneer in the field, Ken Stevens, Bjorn Lindblom, John Ohala). The simultaneous consideration of *phonetic constraints* and *cognitive factors* has led to important advances in understanding the typology of vowel and consonants. In each observation, the phonetician tries to propose an explanation, the most plausible one: anatomical, aerodynamic, acoustical, perceptual or visual explanation. Phonetic explanations should be regarded as hypotheses, and observed tendencies do not have the force of law: two (or more) of the constraints may have a contradictory effect and the outcome cannot be predicted.

It is important to note that a phonological system is the result of a *compromise* between the cognitive tendency favoring the use of a minimum number of features (hence a *symmetry* of the phonological system) and the other constraints (in particular the articulatory and acoustico-perceptual constraints) tending to eliminate combinations of features difficult to realize or distinguish (resulting in an *asymmetry* of the system). Several type of constraints may come into play. For example, the contrast /i — y/ is the most attested case of spreading/rounding contrast, most likeky because both of articulatory ease to round the lips in front vowels and perceptuo-acoustic efficiency. Firstly, it is easy to pronounce /y/. Pronounce an [i] sound with spread lips and then round your lips and project them forward. This is done easily, and you hear [y] (corresponding to the grapheme *u* in French and ü in German). Now notice the difficulty you would have if you do the same gesture of rounding the lips with the low vowel /a/! The action of (spread/rounded) lips is favored *articulatorily* when the mandible is in its high position, and therefore for the high vowels (/i, y, u/). As a result of this articulatory difficulty, languages do not have many oppositions between rounded and unrounded low vowels. French and German contrast round and unrounded front vowels, but have no rounded low vowels. Secondly, it is easy to perceive the contrast between /i/ and /y/. Resonance properties of the vocal tract ensure that the *acoustic* consequences of a lip configuration change are the greatest in the case of front vowels of the /i/ type (namely, /i, e/) (See Fant's nomograms, (Fant 1960a). The constrast between /i/ and /y/ is easy to produce and to perceive.

Today, experimental phonetics and laboratory phonology have moved closer to each other. Theoretical models made by phoneticians to explain the phoneme inventory of systems, i.e. models based on the sound substance, are at least as powerful as the more abstract models offered by phonology (such as the Optimality Theory (Prince & Smolensky 1993).

The convergence of phonetics and phonology has been ongoing for several years by the regular organization of international meetings called *Laboratory*

Phonology. Some differences still remain between the phonetic and phonological approaches. The phonologist is generally guided by a theoretical and deductive approach that determines the hypotheses he wants to verify experimentally. The phonetician is more directly dependent on experimentation: having it in mind at the outset to test his hypotheses by reproducible experiments, he tends to greatly reduce the scope of his research. Moreover, his attention is drawn to the details of the data he collects, which does not provide direct information on language categories but can contribute to the understanding of the many forces acting at any time on the linguistic system and which can, over time, lead to essential sound changes. A centrifugal tendency further pushes the phonetician to look for possible explanations in phylogeny, ontogeny, sociology and anthropology or psychology, while phonology seeks to be closer to the cognitive sciences.

It is therefore more necessary than ever that phoneticians and phonologists understand each other, and this is a constant challenge.

2 Branches of phonetics

The branches of phonetics are essentially articulatory phonetics, acoustic phonetics, auditory phonetics, and neurophonetics, depending on the concerned phase in the speech communication chain, respectively, the speech production apparatus, the acoustic signal, the reception of the signal by the ear and auditory processing, and the brain.

1) *Articulatory phonetics* and orthoepy involve the study of the *correct* pronunciation of words). They are among the oldest branches of linguistics. The Hindu grammarian Panini had already offered a detailed description of the articulation of Sanskrit sounds in the 6th century A.D. in order to set down rules for the *correct* pronunciation of religious texts. It was based on introspection.

   The expansion of the field of phonetics with regard to the types of questions studied beyond its articulatory and orthoepic aspects was largely due to the emergence of new exploratory techniques in the twentieth century. A large number of exploratory techniques were available for studying the *production* aspects of speech, for the study of normal and pathological speech

   - *static palatography* to determine the region of the hard plate contacted by the tongue for the production a given consonant, (Ladefoged & Broadbent 1957),

   - *linguograms* to observe the part of the tongue that is making the contact for the production a given consonant (palatograms and linguograms were often used together),

   - *ElectroPalatoGraphy* (EPG), to study the spatial-temporal course of tongue contacts with the hard palate. (Fletcher et al. 1975),

   - *static X-ray*, (for x-ray data on French (Bothorel et al. 1986), for x-ray data on English, (Perkell 1970),

   - *cine X-ray*, techniques, to study illustrating the position of all speech organs over time during speech, for example: (Delattre & Freeman 1968),

- *x-ray microbeam* techniques (Fujimura et al. 1973),

- *ElectroMagnetic Midsagittal Articulometer* (EMMA) allowing to track the displacement of pellets affixed to various articulators, (Perkell et al. 1992),

- *ultrasound and Magnetic Resonance Imaging* (MRI) (for example, (Baer et al. 1991).

For a review of laboratory techniques for investigating speech articulation and the different imaging techniques (Narayanan & Alwan 1996), (Stone 1997). (Marchal & Cavé 2009).

2) *Acoustic phonetics* really starts with the invention of the invention of *The sound spectrograph*, invented and developed in the 1940s at Bell Labs. The sound spectrograph allows a visual representation of speech, allowing the start of intensive studies of the *acoustic* aspects of sounds. Acoustic phonetics is close to physics and aerodynamics. It studies the acoustic properties of sounds and their relation from the perspective of articulation and perception (see Chapter 5).

3) *Auditory phonetics* and speech perception own much to the invention of the the *Pattern Playback*. The *Pattern Playback* was also developed in the late 1940s, at Haskins Labs It allows to convert the formants seen on a spectrogram back into sound. It had contributed tremendously to the study of the fundamental *perceptual* aspects of the perception of vowels (the formant) and consonants (the role of the formant transition in the identification of the place of articulation). Auditory phonetics is close to physiology of hearing, psycholinguistics, psychology and psychoacoustics. It is concerned with the reception of the acoustic signal by the auditory apparatus and the identification of the linguistic material. Phonetics has greatly contributed to the establishment of audiometric standards (see Chapter 8).

4) *Neurophonetics* is a relatively new and actively expanding discipline. It aims at the understanding of the brain mechanisms underlying speech communication. Medical imaging techniques and evoked potentials now make it possible to compare the levels of activation in different brain areas during speech perception and thus supplement the data provided by the study of language dysfunctions in brain-damaged patients. This is in keeping with the work of the neurologist Paul Broca on corpses in the middle of the 19th century, who discovered the importance of the frontal lobe in speech (Broca 1861). Very recently, a number of *medical neuro-imaging*

techniques include *electroencephalography* (EEG), recording of electrical activity arising from the human brain, and *Functional Magnetic Resonance Imaging* (fMRI) that measure the small changes in blood flow that occur with brain activity. (for a review, (Price 2000). Neurophonetics is at the crossroads of cognitive sciences, neurology and linguistics. Medical data has revealed important differences between individuals and adaptability of cerebral nerve cells, which are organized during the acquisition of the mother tongue and partly reorganized in the event of a brain injury. Despite the individual differences and the adaptability of cerebral nerve cells, which make definitive statements difficult, it appears that the distinctive elements (such as the phonemes and the tones in tonemes languages) necessary for the literal understanding of the successive words of an utterance activate the left hemisphere more strongly whereas the interpretation of emotional prosody would rather be treated in the right hemisphere (just as in music) (Sundberg et al. 1990), (Ziegler 2008), (Stemmer & Lacher 1998), (Guenther 2016).

Three different approaches can be distinguished: *taxonomic*, *experimental* and *applied*, if one excludes phonology, which deals with the architecture of linguistic representations underlying the sound form of language.

1) Until the 19th century, articulatory phonetics was *descriptive* and *taxonomic* in essence. It consisted of describing, representing and classifying the observed facts without seeking explanation.

2) *Experimental phonetics* was born in the middle of the 19th century from the contact between the objective of historical linguistics to explain the reasons for sound changes, on the one hand, and natural sciences such as medicine, physics, botany, anthropology and acoustics, on the other hand. The description of observable facts in the study of language as in other fields of science is only a first stage of research, falling short of the explanatory stage. Abbot Rousselot, founder of experimental phonetics at the end of the 19th century, tried to reproduce the mechanism of phonetic changes in his laboratory. *Experimental phonetics* aims to explain, on the basis of reproducible scientific experiments conducted with more or less sophisticated instruments on a limited set of data (or now also with statistics on very large databases), all the observed manifestations of sounds.

Further readings on experimental phonetics : (Ladefoged 1967), (Lindblom 1986), (Ohala & Jaeger 1986), (Lass 1996), (Lass 2012), (Hayward 2014).

3) Finally, the *applied aspect of phonetics* is ubiquitous and is affirmed by most phoneticians. Applied phonetics include voice technologies such as automatic text-to-speech synthesis, automatic speech recognition, the use of assistive technology tools to help individuals with hearing or speech impairments, and more recently, various applications in the clinical field, for a review (Syrdal et al. 1994).

Several branches within phonetics are commonly distinguished, depending on the domain of applications.

1) General phonetics , like linguistics, looks into universal tendencies in languages, particularly on typological bases and by comparing available data on the acquisition of the mother tongue in different groups and tends to explain the universal tendencies.

2) Historical phonetics and sound changes studies the evolution and classification of languages, and reconstructs the past states of languages through the comparison of attested dialects.

Further readings on historical phonetics : (Pope (1934), (Martinet (1955), (Vaissière (1996), (Joseph & Janda (2003), (Honeybone & Salmons (2015). (Ohala (2017). (Chambers & Schilling (2018).

3) Prosodic studies were developed considerably in the second half of the last century due to a pressing need for the synthesis of speech and the expansion of the field of linguistics. They are currently at the forefront of the international congresses of phonetics. Prosodic studies constitute a vast field, extending from

- *Phonosyntax*, which studies the relations between prosody and syntax, Vaissiére (1971) (Selkirk (1984), Kawaguchi et al. (2006) .

- *Phonostylistics*, which studies the expressive values of the language expressed by the way of saying or pronouncing speech sounds (for example, the voice of a poet, an actor or a politician), (Fónagy 1983), (Léon 1993), (Léon 2009).

- the study of the *identifying* function (i.e. those aspects of the voice characterizing the speaker's dialect, (Clopper & Pisoni 2004), social origin, age (Torre III & Barlow 2009), and personality (Fónagy 1983).

- the study of the *expressive* function (i.e. the expression of personal and impersonal attitudes, (Brown et al. 2015) .

- the study of of the *appellative* function (i.e. a special function serving to arouse certain feelings in the listener, such as compassion as well, (Fónagy 1989).

- *vocal characterology*, which studies the acoustic correlates of individual personality traits as carries by voice (a voice can be felt as gentle, warm, rough or cold (Fónagy 1990).

- and the prosodic aspects of the discourse markers in spontaneous speech (Heeman & Allen 1999).

4) Psychophonetics is concerned, among other things, with the sensations aroused by speech sounds and sound sequences. For example, /i/ evokes the color yellow and /r/ is perceived as more brawling and masculine than /l/ by listeners of various languages (Fónagy 1983). see Chapter 9).

5) Speech therapy phonetics (rehabilitation) treats speech disorders in children or patients who have undergone surgery in the ENT sphere (Ball & Lowry 2008).

6) Clinical phonetics is at the crossroads of linguistics and medicine. It gives priority to the use of tried-and-tested methods of experimental phonetics for the survey of production and perception in pathology. The study of pathological cases of speech has long been a traditional source of phonetic knowledge. Recent medical advances in ENT cancer treatment allow, in some cases, to take into account, beyond the expected survival period, the life quality of patients who have had surgery. In this regard, phonetic knowledge allows doctors to know more about the impact of certain surgical procedures involving the speech organs, and about the quality of voice and speech. Similarly, advances in cochlear implants have clarified the problem of sound coding in the auditory nerve and the learning of sounds. The collaboration between phoneticians and clinicians is also essential to the realization of a large number of experiments carried out with specialized medical equipment, to the construction of data banks of physiological measures that allow the establishment of boundaries between normal and pathological properties, and the evaluation of progress in speech therapy or reconstructive surgery. The exchanges between clinicians and phoneticians have always been very fruitful, based on the sharing of problems, databases, methods and use of technical instruments. Remarkable recent advances in the field of medical imaging (three-dimensional visualization of the speech organs in motion, measurements of magnetic fields induced by the activity of neurons, etc.) have made it possible to broaden the field of phonetics and provide new collaboration with radiologists and neurologists (Ball & Code 1997), (Shriberg et al. 2003), (Windsor et al. 2012),

Further readings on clinical phonetics: (Shriberg et al. 2003), (Ball et al. 2008), (Bauman-Waengler 2012), (Shriberg et al. 2018), (Ball 2021).

Journals on clinical phonetics: Aphasiology, Clinical Linguistics and Phonetics.

Further readings on neurophonetics: (Ziegler 2008), (Stemmer & Lacher 1998), (Stemmer & Whitaker 1998), (Guenther 2016).

Journals on neurolinguistics and phonetics: Acta Neurologica, Brain and Cognition, Brain and Language, Brain Research, Cortex, Hum. Brain Mapping, Nature neuroscience, Neurobiology of Language, NeuroImage.

7) Developmental phonetics , close to psycholinguistics, is concerned with the reactions of the fetus to various sound stimuli, the acquisition processes (perception and production) of the segmental and prosodic characteristics in babies in their mother tongue, and the interference between two languages in the case of bilingual children), in infants, in children and in adults, (Boysson-Bardies 2009).

Journals on child development: Developmental Psychology, Early Childhood Research Quarterly, Infancy, Infant Behavior and Development, Journal of Child Language, Child Development.

8) Didactic phonetics for foreign language learners. It includes setting out standards for the pronunciation of sacred texts since antiquity, second language learning (the International Phonetic Association began as an association of language teachers at the end of the 19th century), (Derwing & Munro 2015), (Doughthy & Lonh 2003).

Journals on second language acquisition: The Journal of Second Language Pronunciation, Studies in Second Language Acquisition).

9) Voice technologies cover scientific research on

- *automatic segmentation and transcription of speech*, (Yu & Deng 2016),
- *automatic text-to-speech synthesis* (i.e. the translation of written text into speech by computer), (Dutoit 1997),
- *automatic speech recognition* (i.e. the translation of spoken language into written sequences of words by computer), (Yu & Deng 2016),

- *spoken man-machine dialogue* , with voice user interfaces, such as Alexa, Cortana, Google Assistant and Siri nowadays, (),
- *automatic speaker identification and forensic phonetics*, using the features of a person's voice to ascertain the speaker's identity and contributing to legal inquiries that require the identification of a voice.

  (Further reading on voice identification: (Jessen 2002), (Jessen 2008), (Lee et al. 2012), (Coulthard et al. 2016), (Visconti 2018),
- *automatic identification of the language*, (House & Neuburg 1977). as well as the identification of some *voice disorders by computer* (Titze & Martin 1998), (Ball et al. 2008), (Damico et al. 2010).

10) Statistical or computational phonetics with the analysis of acoustic-phonetic parameters over a large speech database is expanding significantly. The ever-increasing power of computers, coupled with advances in database storage techniques, and the availability of free segmentation and labelling programs, makes it possible to accumulate large labelled speech corpora, both read and spontaneous. (),

The approaches in the many fields mentioned above were first *knowledge-based*, but *statistics-based*, approaches tend to be more and more used. Knowledge-approaches and statistical approach are complementary. The first teams dedicated to these tasks included close collaboration between engineers and specialists in spoken communication and phoneticians, to understand the relationship between the acoustic characteristics of the individual phonemes and their decoding by the human ear (knowledge-based approach). The studies were done mainly on a small amount of data, in well controlled experiments and in the laboratory. Afterwards, statistical modelling of speech based on training on large amounts of data took precedence over analytical methods in the area of automatic speech recognition and other fields. For example, the use of speech knowledge in automatic speech recognition (ASR) based decoding by eye the distinctive features on spectrogram was extensively used at MIT in the 1980 (Zue 1985). ASR is mainly based on statistical methods. Similarly, concatenative synthesis (concatenating short samples of prerecorded samples derived from natural speech) replaced formant synthesis by rules, which required a great deal of expertise in the field of phonetics (Klatt 1980). In addition, it is through the use of tried-and-tested statistical methods that one can extract from these immense databases the knowledge that reinforces or questions certain conclusions drawn in previous publications. It now plays a leading role in many voice technologies, and has managed to integrate into linguistic theories in terms of probabilistic theories. The functional (statistical) performance of phonemic oppositions and morpho-phonological processes now has a recognized importance in the evolution of languages: less profitable oppositions between two phonemes (that is, those differentiating only a very limited number of words in a language, such as the opposition between /a/ , patte /pat/, ”pawn” and /ɑ/ pâte /pɑ̃t/ ”noodle”, in French) tend to disappear. It is now possible to carry out counts of various types on large labelled databases within a language, or in different languages, or during child language acquisition for the purposes of comparison (Maye et al. 2008), (Pierrehumbert 2003). The absence or insufficiency of metadata (data or information about the data and the linguistic background of the speakers), on one hand, and the lack of standards for annotating the prosodic characteristics of speech, such as speech quality, the grammatical organization in spontaneous speech, modalities, attitudes, emotions, etc., on the other hand, however, limits the potential of *computational phonetics* to go from statistical descriptions to explanations of the variability.

11) Databases collection had become a major topics in phonetic research.

Databases are distributed by the LDC (Language Data Consortium, (Liberman & Cieri 1998), in the United States and by ELRA (European Language Resource Association, (Choukri et al. 1999), in Europe. The LACITO Archiving Program for recordings of glossed and translated rare languages, and the Digital Resource Centers (ADONIS Infrastructure) allow independent laboratories and researchers to freely share their oral data.

In the case of French there are two international projects in progress. The first is called *the Phonology of Contemporary French (PCF): Uses, Varieties and Structures* that aims to construct a transcribed database of samples of French spoken by natives available to everyone (Durand et al. 2005), (Durand et al. 2014). The second, called *The InterPhonology of Contemporary French (IPCF)*, aims to make available corpora of French as spoken by non-native speakers or learners (Detey & Racine 2015),

The documentation of languages that are rare or in danger of disappearing also benefits from new technologies: the archival site of LACITO (LAngues et CIvilisations à Tradition Orale) provides free access to recordings of little-known languages, transcribed on the spot with the help of native speakers, and enriched with a detailed annotation so that they can be accessible to the scientific community on the Internet. In the case of the many

> languages in danger of disappearing in the coming decades, audio data collected by linguists likewise represents a linguistic and cultural heritage, the permanent conservation of which can be ensured through digital techniques. Much work remains to be done.

To conclude, the current state of documentation in phonetics falls short of technical possibilities. Future evolution will undoubtedly make it possible for experienced, as well as junior, researchers to obtain access to complete original data on which the conclusions of the former publications are based. They can thus take an up-to-date look at the linguistic theories and models proposed to them on the basis of data on languages with which they are not familiar. Without access to the original data, particularly complex in case of spontaneous speech, there is a risk of misunderstanding between researchers. Ferdinand de Saussure taught that a linguist should know as many languages as possible. However, due to the increasing specialization of each researcher, only a minority of them concerned with phonetics has a first-hand familiarity with a large number of languages. The quality and the abundance of shared resources are therefore crucial for research in order to be sufficiently open to the variety of languages and to lead to a more and more complete understanding of the phenomena: phonetics should continue to be a science of cumulative nature.

Better training in new phonetic tools for speech-language pathologists, future language teachers and ENT physicians would greatly improve practices of reeducation and learning and would undoubtedly have benefits for basic research. Too few is done in this direction.

Statically-based modeling techniques are more and more able to integrate the knowledge accumulated in the speech science community

The application of phonetic knowledge in the domain of voice technologies, language teaching, and more recently, in the clinical domain has also enriched phonetic research issues. Finally, a considerable factor of recent expansion comes from the willingness of language sciences to approach linguistic phenomena hereafter in their entire cognitive field and to expand the study of language to the uses and behavior of language users from a biological perspective.

As a branch of language sciences, phonetics is at the crossroads of human sciences, life sciences and physical sciences. Phonetic knowledge is essential to audiology, experimental psychology, voice technologies and speech signal processing. The number of professional phoneticians is not increasing but the disciplines dealing with the traditional issues of phonetics are expanding vigorously.

3 The tools of phonetics

Drawing up the system of contrasts between phonemes, tones, or voice qualities (i.e. *the phonological system* of a language) used to distinguish between words is always the first necessary step in the in-depth analysis of a language. Methods of survey and analysis for decoding the abstract units of the linguistic code are basically the same regardless of whether you are studying a language as yet unexplored or a regional variety of a language the other varieties of which have thoroughly been explored (Martinet (1956). Sometimes it takes months or years to establish the phonological system of a language that has never been studied before, and phoneme inventories may vary from one regional variety to another or depend on the age. For example, as previously mentioned, the exact number of vowels is not the same in the north and south of France, nor is it for young and elderly speakers: the distinction between /ɛ̃/ and /œ̃/ still exit in south France; young speakers do not distinguish anymore /a/ and /ɑ/.

The most important tool for representing the phonological system of a language is the *The International Phonetic Alphabet* (IPA). IPA was devised by Otto Jespersen in 1886 for phonological purpose. IPA symbols allow to describe the phonological system of a language using a fixed number of symbols, each symbol representing a single phoneme. The IPA is a means that remains perfectible but has the decisive advantage of being quite uniformly used for describing the phonological system of most of the studied languages, thereby facilitating scientific research and collaboration. Shortly afterwards it was conceived as a notation system by a group of language teachers to meet the need for phonetic transcription within the context of language learning. Their aim was to be describe the way of pronouncing the foreign words by using written symbols as accurate as possible. The IPA is therefore used both as a phonological transcription system for describing the system of contrasts between phonemes in a given language, and for transcribing phonetically the actual realization of the speech sound (eventually with the use of diacritics), (Ladefoged (1990), (Association (1999).

The two basic principles of the use of the IPA as a phonological transcription system are as follows (Association (1999)).

1. There should be a strict distinction between a distinctive, abstract sound (i.e. a *phoneme*) and its different phonetic realizations (the so-called variants or *allophones*). A phoneme in a language is represented by a single symbol between two slashes, / /, its realisations between two its realisations by square brackets, [ ]. Let's give some examples.

   a) The standard French r sound, transcribed by the symbol /ʁ/, representing the uvular voiced fricative [ʁ]. /ʁ/ is realized very differently depending on the speaker and on the context. The *Burgundian r* (realized with the tip of the tongue vibrating against the alveolar ridge), [r], the velarized realization of /ʁ/ in *rourou*, [ʁˠuʁˠu], the palatalized realization in *riri* [ʁʲiʁʲi], the realizations with more or less friction noise or no noise at all, as well as the unvoiced versions of /ʁ/ as in *tra* [tʁ̥a] or as the approximant version of /ʁ/ (such as in *rara* (the narrowing of the vocal tract is not enough to produce a turbulent airstream) are phonologically transcribed by a single phoneme symbol /ʁ/ in French because their substitution for each other is not distinctive — it cannot be used to distinguish between two French words.

   b) Similarly, [t], [tʰ], [t̬], [ɾ], [ʔ] are variants of the same phoneme /t/ in English, but they represent the two phonemes /t/ and /tʰ/ in Hindi.

   c) /t/ is dental on French, and alveolar in English, but only one symbol, /t/, is used. Malayalam uses dental and alveolar /t/ contrastively (Jongman et al. (1985)).

   d) In Japanese the syllable /sa si su se so/ and /ta ti tu te to/ are respectively pronounced [sa ɕi sɯ se so] and [ta tɕ tsɯ te to], i.e. /si/ > [ɕi] /ti/ > [t͡ɕi] /tu/ > [tsɯ]. [ɕ] is a contextual, predictable, allophone of /s/ in Japanese (/s/ is always pronounced [ɕ] when followed by [i]) (the sequence [si] does not exist phonetically in that language) [t͡i] and are contextual, predictable, allophones of /t/ (the sequence [ti] does not exist phonetically in that language)

   The exact symbol used for a phonemic (broad, phonological) transcription in a language is, up to a certain extent, a matter of choice made by the transcriber or a matter of convention. For example, in Hungarian, there is a contrast between short and long vowels, but /e/ and /ɛ/ are not two distinctive phonemes, the short vowel is often usually represented as /ɛ/ and the long vowel as /eː/, reflecting both the change of timber and the change of duration. The symbols for describing the vocalic system of Japanese are either /i, e, a, o, u/ or /i, ɛ, a, o, ɯ/ʊ/. The same set of symbols /a/, /e/, /i/

and /u/ are extensively used to describe vowels in five-vowel languages for the sake of simplicity although the vowel timbers may greatly differ from one language or dialect to another. For illustration of the variations in the realisations of the ”same” phoneme in different languages, (Disner (1983)).

2. If the *same* symbol is used in two different languages for representing a phoneme, it has to correspond to sounds of *identical or very similar timber* and to a *same* or *similar articulatory configuration.* of the vocal tract. This principle is not always applied rigorously because the concern for typographical simplicity leads to some compromises: /t/ in English may be realized as glottal stop, and a glottal stop does not correspond to a similar configuration of the vocal tract, See chapter 7).

   The choice of the exact symbols do not really matter for a phonemic transcription.

Speech can be transcribed phonetically at different levels of detail and accuracy [1].

1. A *phonemic* (abstract, phonological) transcription disregards all allophonic differences and use only the fixed set of symbols corresponding to the phonemes in the particular language (all /t/ variants are transcribed as /t/). It disregards all of the alternations in pronunciation that are predictable by phonological rules, as mentioned above.

2. A *broad transcription*, as generally found in the dictionaries, indicates the most usual way of pronouncing the word in the dialect; The words ”potato”, ”butter” and ”schedule” are respectively transcribed as [pəˈteɪtəʊ], [ˈbʌtə] and [ˈʃɛdjuːl] in British English and as [pəˈteɪtoʊ], [ˈbʌtɚ] and [ˈskɛʤʊl] in American English. The differences in the broad transcription reflect the expected differences in pronunciation between the two dialects of English.

3. A *narrow phonetic transcription* encodes more information about the precise phonetic details of the allophones in the utterance, such as the effects of coarticulation (contextual nasalisation, rounding, voicing or devoicing, etc.) and make use of the diacritics available in the IPA. /t/ may be transcribed as non aspirated [t] (star [’stɑr]), as aspirated [t^h] (tie [’taɪ]), as a

[1] Alexander Ellis (1814-90) introduced the distinction between different levels of phonetic transcription, *The Alphabet of Nature*, 1845.

> tap [ɾ] (writer or rider [raɪɾɚr], American English), as a nasal [n] (twenty [ˈtwɛnti]), as an affricate [tʃ] (two [ˈtʃʉ]) and glottal stop[ʔ] (bit [’bɪʔ]).
>
> *Sixty-seven diacritics* (a glyph added to a symbol) may be used to show subtle variations. For example, the diacritic ^{w} in [t^{w}] indicates the lip-rounded realization of the consonant /t/, and the diacritic + in [u̟] indicates the advanced realization of the vowel /u/ as in *toute* [t^{w}u̟ t^{w}]. annotating prosodic events such as tones, lengthening, phonation quality, etc. The number of diacritics used for a phonetic transcription is a matter of choice made by the transcriber, his training and of the purpose of the transcription.

A narrow phonetic transcription based only on listening is often difficult to make even with a highly trained ear.

Firstly, few phoneticians master the whole set of the available phonetic symbols of the IPA. Furthermore, the phonetic transcriptions of the same speech material made by two different trained transcribers may not be identical. Their way of labelling depends in part on the contrastive sounds and the patterns of alternation on their native language and on the languages they master. It also depends on the extension of their training in phonetic transcription.

Secondly, the size of the time window of the acoustic signal (a phone, a syllable, a word, a whole sentence) corresponding to the speech portion to be transcribes is crucial for the phonetic transcription of the segment. For example, when a native French speaker hears the portion of the signal corresponding to the entire word *rire* /ʁiʁ/ ‘laugh”, he will transcribe it as [ʁiʁ] without hesitation. If the portion of the signal corresponding to the (uvularized) realization of /i/ in /ʁiʁ/ is extracted, he will hesitate between [i] and [e] timbers. Due to coarticulation with the two flanking back consonant /ʁ/, the tongue during the production of /i/ is somehow lowered and placed in a more backed position than expected: the resulting sound is at mid perceptual distance between a hyperarticulated [i] and a standard [e]. If he is native of French, but knows how to transcribe English in the IPA, and is familiar therefore with the timbre [ɪ], he will probably propose [ɪ], like in “bit” /bɪt/ and indicated that it is lengthened [ɪː].

Thirdly, if the researcher also uses a visual representation of what has be to transcribed, such as a spectrogram, he may be influenced but what he sees. This point is particularly important for transcribing the pitch contours of an utterance: there is often a discrepancy between what is heard and what is seen.

The *ear* remains the main working instrument of the phonetician for transcribing speech and is the only ultimate judge. A single acoustic or articulatory fact no matter how in-depth and thorough it may be, does not allow for definitive conclusions to be drawn on the way the sound is perceived by an human ear or used by the auditory system. Perceptual experiments with native speakers of the language are always necessary to conclude about the use of such and such acoustic parameters in speech perception.

Firstly, subtle variations observed acoustically, however regular, may not be consciously perceived. If they are not consciously perceived, they may however play a role in the speed of identifying phonemes (i.e. acceleration of *reaction times* For example, when listeners were faster and more accurate for identifying the vowels /i/, /a/, and /u/ in cross-spliced friction noises across tokens of Isal, Isul, Isal, and Isul, when the frication noises provided accurate anticipatory information for the vowels than when they provided misleading information (Martin et al. (1980), Whalen (1984)).

Secondly, if the subtle variations are perceived, they may carry several types of information for native listeners. For example, in French, the use of a more palatalized variant of a consonant may add a *nuance of kindness or tenderness* to what is expressed Fónagy et al. (1983); it may be a cue for identifying a *regional accent* de Mareüil (2010) or for identifying the *sociocultural background* of the speaker Martinet (1945) or the sign of a *speech disorder* Schoentgen (2006).

Modern techniques may complement the judgment of a listener on the phonetic aspects of an utterance and make it possible to carry out highly elaborate perception experiments on phoneme identification and discrimination with reaction time measurements or magnetic field measurements induced by the activity of neurons. (Näätänen et al. (1997). They even make it possible to test sound discrimination in a sleeping baby by EGG (expounded upon later in the book). By the age of one year, the brain react differently for native phonemes and for non-native phonemes (Cheour et al. (1998).

The *spectrographic display*, the *measured formant frequency values* (as we will discuss later) and *listening* to each individual sound segment of the text to be transcribed have become indispensable for an accurate narrow phonetic transcriptions. For example, lip rounding (and subsequent darkening of the timber) during the production of the initial consonant /s/ in the French word *structure* as compared to the word *stricture* in not perceived by the most trained ear because it is *expected*. The audible darkening of the timber in /s/ in the word *structure* clearly stands out on a spectrogram as a noticeable lowering (of about 1000 Hz!) of the resonance frequencies during the fricative portion. A separate listening of the friction noise corresponding to the two /s/ variants, confirms the differences in timber (Benguerel & Cowan (1974). The degree of protrusion or spreading of the lips can also be measured on images or videos of the face and can be mathematically related to the degree of lowering in the resonance frequencies through equations, Fant (1960b).

The devil is in the details, thus rendering an account of very subtle synchronic variations may shed light on the origins of well-attested diachronic changes: sound changes emerge from synchronic variations Ohala (2011). see for example the resemblance between diachronic changes and synchronic variations in the passage from Latin to Modern French, Vaissière (1996).

The three types of characteristics of speech sounds including perceptual, articulatory and acoustic cases are intimately related to each other and understanding their relationship is essential to new discoveries about speech functioning. Ideally, the description of a phenomenon should include its articulatory, acoustic, and perceptual aspects and explain their relationship.

A well-founded narrow phonetic transcription based on acoustico-perceptual parameters should highlight the system of acoustic cues that allow phonemic oppositions. It should also be capable of accounting for the acoustic differences between the phonetic realizations of a single phoneme (e.g. /i/ or /s/, etc.) in different languages or language varieties and reflect the acoustic nuances that convey information for native speakers within a language variety.

One important remark concerns the important of the third formant. The first two measured formants on spectrograms have been unanimously recognized as the most important acoustic cues for vowel identification, and this statement is true. Vowels may be approximated to a fair degree of phonetic quality by two formants. Delattre and his colleagues, Delattre et al. (1952) demonstrated using the pattern playback that two formants were sufficient to synthesize the color of all French oral vowels. The second formant used for the synthesis of /i/ (2900 Hz) in their experiment was mistakenly considered as the real F_2 of /i/; in reality, it was the third formant. This often cited study and others contribute to underline the role of the third formant. The role of the third formant has been unfortunately very neglected in the literature. The formant synthesis (such as Dennis Klatt's text-to-speech system, Klatt (1980), in which the computer reproduces speech from the formant frequencies indicated by the experimenter (see Chapter 5), demonstrates the necessity of taking into account the third formant (abbreviated F_3), at least in languages that display a phonological contrast between rounded and unrounded front vowels (such as in French, Swedish and German). The French vowels /i/ and /y/ can be produced or synthesized with identical values of F_1 and F_2, and be distinguished only through their F_3 frequency value. It is noteworthy that the information provided by F_3 is also necessary to account for the timber difference in the realization of the French /i/ (identical to Swedish /i/) and the British /i/ (with lower F_3, Vaissière (2011). Similarly, the timber of the North American English [ɚ] and [ʌ] can be synthesized with identical F_1 and F_2 frequencies, with a F_3 higher than 2000Hz for [ʌ] and much under 2000 Hz for [ɚ]. F_3 and higher formants have a perceptual weight at for the vowels with mid and high for not grave vowels F_2, Carlson et al. (1974).

The choice of acoustico-perceptual parameters to be investigated is often influenced by the researcher's mother tongue, and the role of F_3 is rarely acknowledged in English phonetics books simply, and unfortunately, maybe because English has no opposition between spread and rounded vowels ?

The same timber can be produced by different configurations of the vocal tract. The timbers produced through articulatory synthesis (like that of Shinji Maeda, Maeda (1982) , based either on indications given to a computer on the position of the speech organs, or on a simplified diagram of the vocal tract, show the importance of compensation mechanisms (called *trading relations*) between the lips and the tongue, and between the tongue and the jaw, Maeda (1990). For example, the front cavity can be lengthened (and therefore the formants sensitive to the front cavity configuration are lowered) by lip protrusion and/or tongue backing, in which case both gestures can compensate for each other, or reinforce each other when used together (lip compression has the same effect than lip protrusion: it lowers the formant frequencies due to the front cavities). The two distinctive features related to roundness and backness should not be considered separately in any theory because their acoustic effects on the frequency value of the formants due to the front cavities are acoustically similar. The two distinctive features are not orthogonal: lip-rounding and tongue-backing are to necessary conditions for the vowels to be perceived as grave (such as /u/, /o/, /ɔ/). The targeted movement of the lips makes it possible to compensate to a great extent for the position of the tongue. A rounded palatal vowel (top) is acoustically very close to a more back non-rounded vowel (bottom), which, if rounded, results in the French /u/ sound (articulatory model). The formant frequencies are similar.

Similar, the consonant /ɹ/ and vowel /ɚ/ can be produced by different tongue postures: retroflex or bunched, rounded or not rounded; listeners do not perceived the difference as so far the third formant is lowered under 1800 Hz and cannot recover the tongue and the lip configuration, Delattre & Freeman (1968)).

Jakobson expressed the idea that speech sounds should be ultimately transcribed on the basis of their auditory characteristics (such as grave or acute), rather than their articulatory (e.g. labial, dorsal, back or front) and/or acoustic characteristics (e.g. strident, low of high F2). This suggests with a new type of

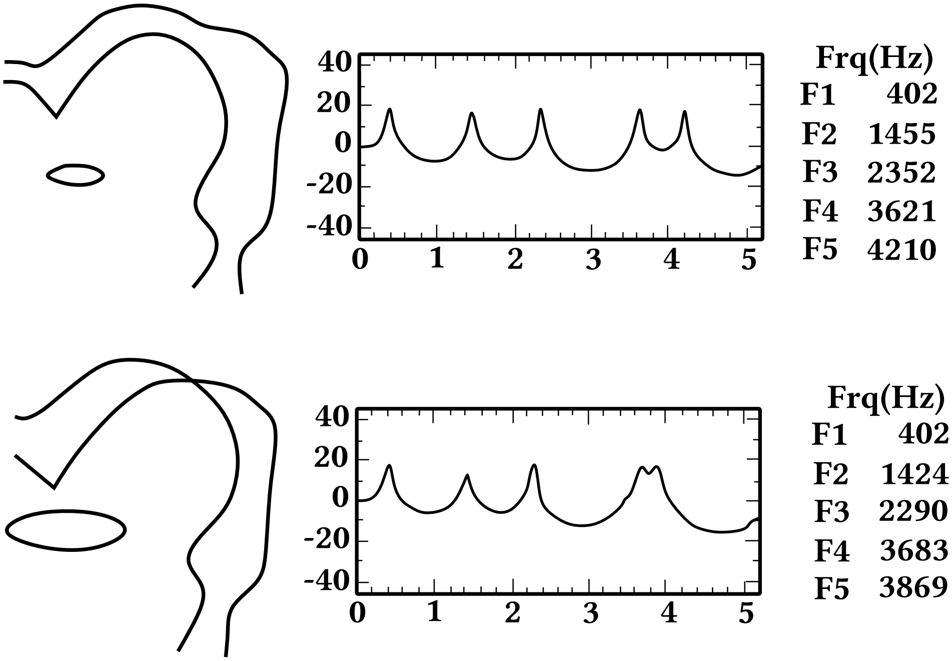


Figure 3.1: Articulatory compensation

phonetic transcription based uniformly on the relevant perceptuo-acoustic properties of the speech signal[2]

An adequate study of the spoken side of a language can only be done on the basis of knowledge of its various dimensions: phonetic, phonological, morphological and syntactic. An high-performance analysis of the phonemic systems of a language, including its vowel and consonant systems, can certainly be carried out through an in-depth survey of the lexicon by an investigator whose mother tongue is different from the one under analysis and who does not master the target language. He has at his disposal a set of methods well established over time by phonology, and the phonemic systems of a language established by a non-native researcher can be trusted, Ladefoged (2003).

On the other hand, it is risky to launch into the description of the prosodic system of a non-mother tongue language: an acoustic prosodic detail can carry information for the native speaker of the language and escape the non-native researcher's attention. One should be very careful with prosodic descriptions done by researchers who do not speak fluently the language they describe and

[2] The system in question is being developed by Vaissiére (2007))

don't master its morphological and syntactic aspects. It is necessary to consider the scope of any observed prosodic fact within the paradigm of all of the processes offered by the language (morphological, syntactic, etc.). For example, the utterance modality (declarative, interrogative and exclamatory) is signalled by prosodic markers and/or morphosyntactic markers, depending on the language. In French, for example, a rising contour on the final segment on the last syllable of the yes–no question is the sole and mandatory indicator of interrogative form for the sentence with declarative syntax *tu viens ?* [ɛskəty vjɛ] "You are coming?". A rising contour is not indispensable for *viens-tu ?* [vjɛ ty] "Are you coming?" and *Est-ce que tu viens ?* "Do you come?" because the interrogative modality is already marked by another means (word order and the use of specific morphemes). A rising intonation in two latter cases where it is not strictly necessary adds a nuance of politeness or kindness, Fónagy (1982).

Information from fieldwork and from laboratory are *complementary*. Since speech is a complex phenomenon, it is necessary to conceive each phenomenon from a number of angles. Field studies provide first-hand data for the actual uses of natural language and allow one to capture or verify phenomena that occur in the real world. Each observed phenomenon from fieldwork could be further investigated thoroughly with a corpus designed for that purpose in a laboratory, to disentangle the effects of different factors.

Phonetic investigations, done in the fields or in the laboratories continuously embraced technological innovation.

With the arrival of the big data era, speech data collection is now much easier than before. However, this does not at all exempt the investigator from data collection training, Ladefoged (2003).

Technical progress makes it possible to collect several types of acoustic, articulatory, visual and physiological data at the same time, and to use sophisticated non-invasive instrumentation under well-controlled conditions that facilitate the interpretation of phenomena by providing, for example, direct information on the configuration of articulators. It is only recently that phonetic laboratories have had at their disposal technologies to capture movements of markers placed on the speaker's face or on articulators, and other non-invasive methods (ultrasound, Stone, 2005, etc.) of high precision, Stone (1997))

The performance of recording devices (e.g. audio recorders, video cameras, electroglottographs, electroencephalographs, scanners, flow and pressure sensors) is constantly improving and their *miniaturization* makes it possible to use some of them outside of the laboratory, in the field. Audio recordings can now be easily complemented in the field by aerodynamic data (air flow and pressure), as well as by palatographic, glottographic, videographic and ultrasound data.

Certain data concerning speech production can only be collected in a hospital because they require the presence of clinicians and heavy equipment generally used by clinicians as a) with *electromyographic* measurements (which study the electrical *activity* of nerves and muscles related to speech production), b) a *cineradiographic* investigation, or a *fibroscopic* investigation of the speech organs or the vocal tract, using a fibroscope inserted inside the vocal tract. *Magnetic resonance imaging* (MRI), *transillumination of the larynx*, which transmits the images of the speech organs, c) *functional cerebral imaging*, *electroencephalography* (EEG) and *magnetoencephalography* (MEG) are also be very useful and are increasingly used in phonetic studies in close collaboration between specialists of different fields. All the data, whether collected in the field, in a phonetic laboratory, or in a hospital allows rapid advances in our knowledge of the (dys)function of the spoken side of the language.

Tools are constantly evolving. The old tools available to researchers in experimental phonetics are being improved and new tools are being launched. We may mention, for example, using the EPGG (external electrophotography)[3] to estimate the glottis size in a non-invasive way. The EPGG is used by speech acousticians at the phonetic laboratory of the Université Sorbonne Nouvelle, Paris.

Whatever the type of data, it is rare nowadays to collect and process data without using *software* tools. Storing and processing data has become increasingly easy, and computerized database collaboration and statistics have become essential tools for phonetic research. Computer open-source software packages for the scientific analysis of speech (such as Praat, Boersma (2011) and statistics (such as R) are freely available.

Finally, it should be noted that the Internet allows individual researchers to keep up to date on databases and publications, to listen to the sounds made available by researchers all over the world, and carry out an acoustic and perceptual study on one's own computer of the same sounds.

We conclude this chapter with one important point. The scope of the experimental results obtained or the statistics from large databases should always be relativized. The type of corpus studied (narratives, descriptions of images, spontaneous dialogues, single words, read texts), the choice of speakers or listeners, the recording conditions (context, instructions provided to the speakers), the inaccessibility of metadata that would be useful to explain sound variations and the theoritical beliefs of the observer, have an impact on the results obtained. Great caution, often not observed in the literature, is required before any generalization. Sharing the data bases becomes essential.

[3] Inventors: Kiyoshi Honda and Shinji Maeda, in Paris, 2009.

4 Speech organs

The study of the function of speech organs belongs to the field of articulatory phonetics, which is the oldest branch of phonetics. By the end of the 19th century introspection was supplemented by static palatography[1], Durand (1930). Studies on the dynamic phenomena of coarticulation truly began in the 20th century with the provision of a number of inventions in the domain of imaging techniques: *radiography* (1895), *electromyography* (1941), *cineradiography* (1954), *dynamic palatography* (1960), *aerodynamic measurements*, the *X-ray Microbeam System* (created in the 1970s by Osamu Fujimura, Kiritani et al. (1975), and *electromagnetic articulatograph*. More recently *ultrafast cameras* placed inside the vocal tract with the help of a fibroscope and three-dimensional MRI provide valuable information on the position and activity of articulators. (see a review on laboratory techniques for investigating speech articulation by Stone (1997). Finally, *articulatory synthesis* alone provides a link between articulation, acoustics and perception, and modeling has become the main source of progress in the field of articulatory phonetics, Kröger & Birkholz (2009).

Humans produce speech with organs for which the main function is not linguistic. Figure 4.1 illustrates the sagittal section of the face and neck, including the main organs involved in speech production: the larynx, the tongue, the velum and the lips.

To *produce sounds* they use the lungs, the larynx, the tongue, the lips and the soft palate.

- The primary function of the *lungs* is respiratory (i.e. body oxygenation).

- The primary function of the *larynx* is the protection of the respiratory tract in humans as well as in animals.

- The main function of the *tongue* is to participate in chewing and swallowing.

The monkey has a comparable morphology to that of the human, but it does not *speak*. Phylogenetically, the emergence of the creative faculty of language in man certainly has to do with the increase of his cognitive abilities and the cerebral areas of Broca and Wernicke, and not to the configuration of his vocal

[1] See the early works of Pierre-Jean Rousselot and Marguerite Durand.

tract, even if the low vertical position of the larynx in the adult man greatly facilitates the mobility of the tongue. In the human baby and in some mammals, a high larynx makes it possible to breathe and drink simultaneously, but it limits the tongue's movements. But a higher position of the larynx would not hinder the development of communication by speech-like sounds made by the vocal tract in mammals: articulatory synthesis shows that the maximal vowel space of newborn infants is potentially (at least) the same as adults, Boë (1999).

The main difference between the monkey and the human is the higher degree of *cortex* development in the latter, Darwin's (1859). It has proved impossible to teach human speech to a monkey. Some chimpanzees (with whom we share 99% of our genetic material!) are able to learn the meanings of some 150 words but do not spontaneously combine those words to make new sentences, Aboitiz (2017).

The act of enunciation can be divided into several phases, called the speech chain:

- a *psychic* phase, in the mind of the speaker, which is the phase of the intent to transmit information ;
- a *linguistic* phase with the selection in the mental lexicon of the words corresponding to the message to be transmitted, the arrangement of these words according to the rules of syntax, and the choice of a prosody appropriate to the overall intent of the speaker's message;
- a *physiological* phase with a set of commands to the relevant muscles of the articulators, the lungs, the larynx, the tongue, the lips and the soft palate, which results in the generation of sound; and an
- an *aero-acoustic* phase. The acoustic wave produced by the speaker is transmitted through the air and causes the listener's eardrum to vibrate; the message is analyzed with
- a *physiological* phase at the level of the ear , which transforms into electrical activity in the auditory nerve; then
- a *linguistic* phase in the higher center of the brain to recover the intended message and finally
- a *psychic* phase, consisting of the interpretation of the message by the listener (Further readings on the speech chain: Denes et al. (1993), Levelt (1989).

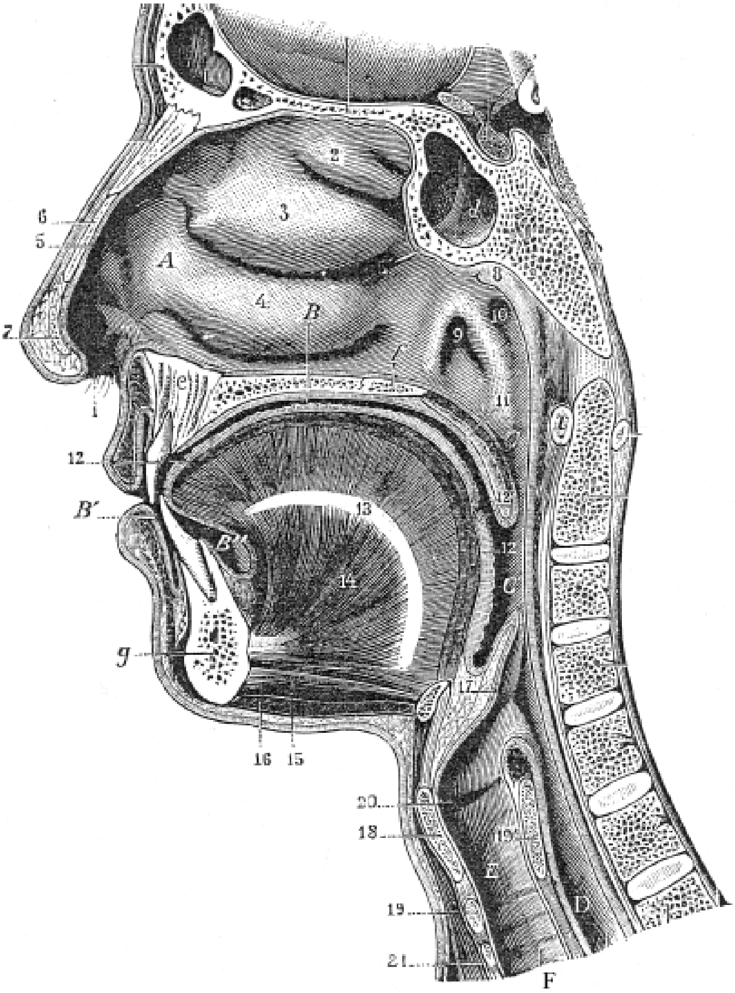


Figure 4.1: Sagittal section of the face and neck, including the main organs involved in speech production. Plate of Testut (1889: *Traité d'anatomie humaine*) used by Abbot Rousselot in his book *Principes de phonétique expérimentale* (1897-1908) A. Right nasal cavity; B. oral cavity; B'. vestibule; B". sublingual region; C. nasal pharynx; C'. oral pharynx; D. esophagus; E. larynx; F. trachea 1. right nostril; 2. superior nasal concha; 3. middle nasal concha; 4. inferior nasal concha; 5, 5'. mucous membrane of the nasal cavity; 6. lateral cartilage of the nose; 7. cartilage of the nose wing; 8. pharyngeal tonsil; 9. pharyngeal opening of the auditory tube; 10. pharyngeal recess (also called fossa of Rosenmüller); 11. soft palate (also called velum) and the uvula; 12. lingual mucosa; 12'. Eoramen caecum 13. fibrous nucleus of the tongue; 14. genioglossus; 15. geniohyoid muscle; 16. mylohyoid muscle; 17. epiglottis; 18. thyroid cartilage; 19, 19'. cricoid cartilage; 20. larynx ventricle; 21. the first tracheal ring.

Speech production involves three main processes: *breathing*, *phonation*, and *articulation* (see Figure 4.2).

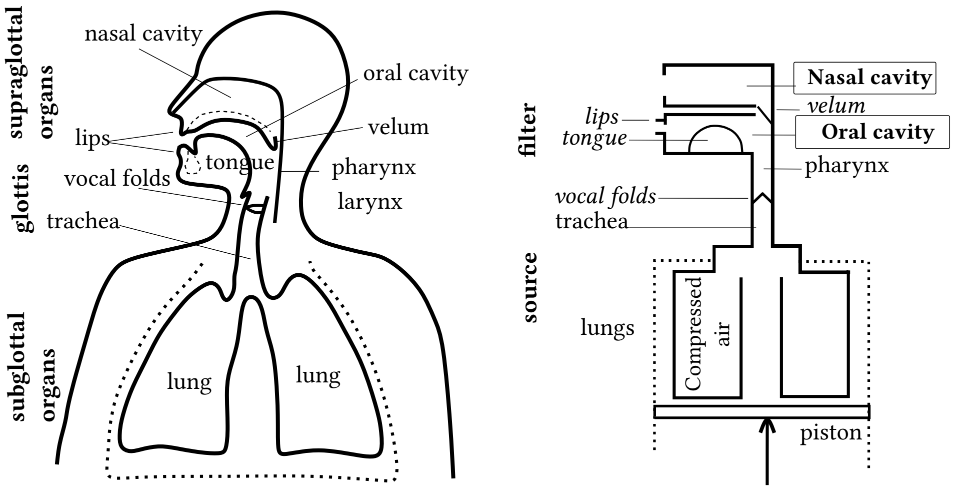


Figure 4.2: The schematic representation of the so-called *speech organs*

The so-called *speech organs* are generally grouped into three types according to their involvement into one of three processes:

- at the subglottal level, the *respiratory muscles* that create the egressive airflow necessary for phonation and production of egressive noises (as in fricative and stop consonants);
- at the glottal level, the *phonatory organs* that make the laryngeal buzzing and
- at the supraglottal level, the *articulatory organs* that filter this buzzing (i.e. the source signal) and produce the different successive sounds (see Figure 4.2).

1. The first phase is the production of the source.

   The subglottal component (lungs, bronchi, trachea and respiratory organs) acts as a wind tunnel. During normal breathing, inspiration and exhalation are of similar duration (respectively, 40% and 60%). When the speaker intends to speak, he inhales a greater volume of air in a shorter time than for normal breathing. The exhalation, during which he emits sounds, will often be ten times longer than inhalation. The movement of the rib cage and diaphragm compresses the air of the lungs like the piston of a bicycle pump, thus creating a very high subglottal pressure necessary for the exhalation of an air stream through the vocal folds. The active muscular forces (of the rib cage, diaphragm and abdomen), combined with passive elastic forces (the elastic property of the tissues), tend to maintain a relatively constant high subglottal pressure, between 6 and 10 cm H_2O. The subglottal pressure sometimes decreases slightly during the production of an utterance. It may sharply increase locally in the case of emphasis, or globally in the case of shouting. The average airflow rate during speech varies from 100 to 300 ml of air per second. The most efficient sounds in terms of total air consumption are voiced stops (50 ml) and vowels, and then voiced fricatives (75 ml). Voiceless stops consume 80 ml and voiceless fricatives 100 ml, see Lass (2012).

2. The second phase is phonation.

   Phonation transforms the airflow coming out of the lungs into a buzz created by the quasi-periodic vibrations of the adducted vocal folds. The air exhaled from the lungs passes through the trachea and reaches the larynx where the vocal folds are located. Figure 4.3 shows the larynx, which forms the upper end of the trachea, and presents different glottis configurations (discussed below).

   The larynx is protuberant in adult men (called *the Adam's apple*. The *vocal folds* (improperly called *vocal cords* in a false analogy with musical instruments) are located in the larynx. They are two vibrating muscles covered with a mucous membrane, which are inserted between the thyroid cartilage, which protects them, and two mobile cartilages (i.e. the arytenoids), which make it possible to modify their length and spacing.

   The space between the two vocal folds is called the *glottis*.

   - If the vocal folds are separated (i.e. if the glottis is open), the continuous airflow passes through freely (as in the production of voiceless sounds [p, t, k, f, s, ʃ] and during breathing.

   - If the vocal folds are strongly tightened (the glottis is closed), the air is blocked (as in the production of a glottal stop).

   - If they are softly joined (as is the case with most speech sounds, which are voiced sounds), the airflow causes them to vibrate if the transglottal pressure is high enough. The vocal folds' vibration cuts the airflow into a

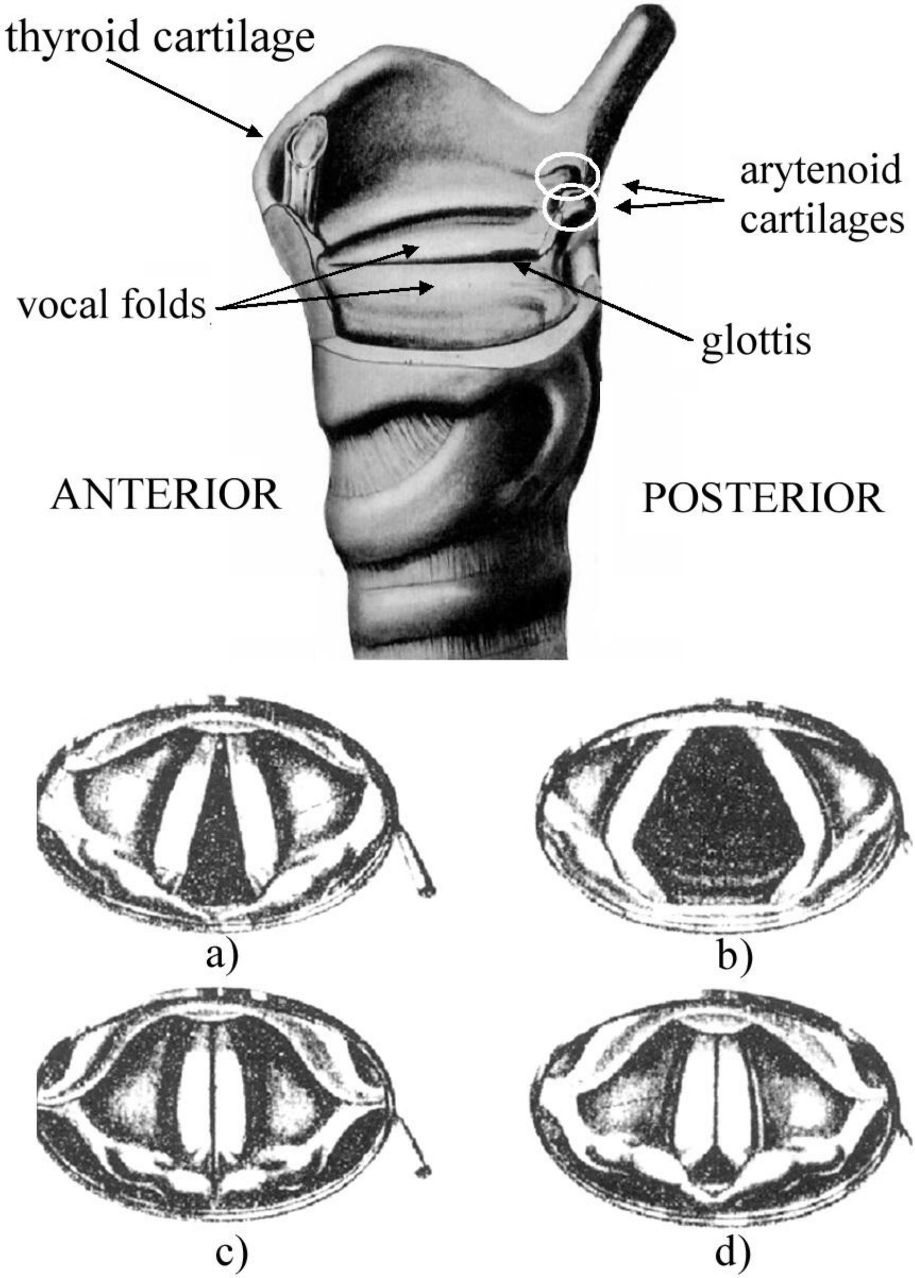


Figure 4.3: At the top: the upper part of the trachea, larynx, vocal folds, glottis and arytenoids. At the bottom: vocal folds positions during a) breathing, b) deep inhalation, c) phonation and d) whispering (after Farnsworth, 1940, top and Pernkopf, 1952, bottom)

succession of puffs of air creating a buzzing, which is roughly the same for all voiced sounds. The vocal folds in this case play the role of an oscillator.

Phylogenetically, vocal folds provide a sphincter function that protects the respiratory tract from the descent of food into the lungs and also allows pressurization of the lungs in the event of a physical effort, and the size of the glottis helps to regulate breathing. The vocal folds are 3 mm long in the newborn, 10 mm at puberty, and increase by 5-10 mm in adult men and 3-5 mm in adult women.

Most of the sounds used in speech are *voiced*, i.e. produced with the intervention of vibrating vocal folds. By placing the palm of your hand on your throat and saying [a], [z] and [s], you will feel the vibrations for the first two sounds, which are voiced sounds, and notice the absence of vibrations for the last one, which is a voiceless sound. If you repeat this test for all French phonemes, you will notice that most of the sounds are accompanied by a vibration except [p, t, k, f, s, ʃ], which are voiceless sounds.

The *subglottal pressure* is about 8-10 cm H_2O during speech, higher than the intraoral pressure and the transglottal pressure is therefore positive. The positive difference between subglottal pressure and intraoral pressure is one of the necessary conditions for the vocal folds to vibrate.

In order to be set into vibration, the vocal folds are joined through a pivot motion of the arytenoid cartilages. A transglottal pressure of 3 to 5 cm H_2O is enough to make them vibrate, and a pressure of 1 to 2 cm H_2O to keep them vibrating. The occlusion in voiced stop consonants /b, d, g/, the constriction in the vocal tract for voiced fricatives /v, z, ʒ/, or the realization of very high vowels /i, y, u/ increases intraoral pressure, and diminishes transglottal pressure, and consequently impedes or delays the vibration of the vocal folds, hence the natural tendency to devoice high vowels — obligatory (phonologized phenomenon) in Japanese when the vowel /i/ or /u/ occurs between two voiceless consonants. The delay of voicing contributes to the affrication of the dental stops that precede /i/ in Canadian French: *ta pt*^{s}*ite voit*^{s}*ure*, but not in the case of /u/.

At the start of the vibration cycle, the vocal folds are brought together softly, preventing the air from escaping *(1)*. With the thrust of the airflow, the pressure under the closed glottis increases, the vocal folds are pushed apart, their contact area diminishes, and they end up separating and the air escapes *(2)*. The escaping air creates a low-pressure zone between the two

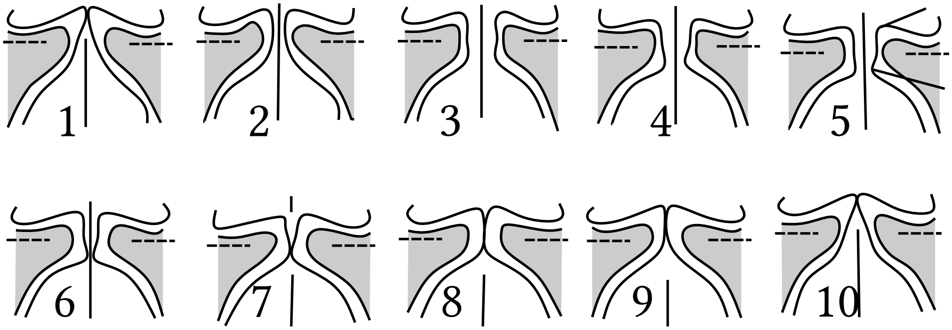


Figure 4.4: . — A cycle of viration of the vocal folds (According to Hirano, 1981)

vocal folds and their myoelasticity pulls them back together from the bottom *(3-6)*. They then snap shut like a door that slams because of a current of moving air *(6)*; the airway is blocked, the air pressure under the vocal folds increases again *(6-10)* and finally blows them apart so that the vibration cycle keeps repeating itself. The force of the closure guarantees an efficient voice and strengthens the amplitude of harmonics in the medium and high frequency range. The quality of the closure of the vocal folds determines the phonation quality, Gordon & Ladefoged (2001).

The average frequency of vibration of the vocal folds depends on the individual, in particular his vocal fold mass, which is related to age and sex. The larger the vocal folds, the slower their vibrational rhythm. They vibrate on average 120 times per second in adult men (120 Hz), 240 times in adult women (240 Hz), 350 times in children (350 Hz), and 400 times or more in newborns (400 Hz). The average frequency of the vocal folds' vibration in an individual changes over the course of his life With age, the male voice becomes grave and the female voice acute.

A speaker can *intentionally* increase or decrease the vibration frequency of his vocal folds during the production of vowels and sonorants up to a certain range. In daily conversation, the effort to modify the vibrational rhythm of the vocal folds is made essentially at the level of the larynx. The vibrational rhythm is controlled by the stiffness of the vocal folds: the stiffer the vocal folds, the higher the frequency of vibrations. The stiffness of the vocal folds is increased when stretched out by a movement of the arytenoids. There is a second way of increasing the vibrational rhythm, which consists of *increasing the articulatory effort* by a greater compression of the lungs, used in the employment of emphasis, shouting, and in certain pathologies. The resulting increase in subglottal pressure increases the frequency of the vocal folds' vibration and the amplitude of the vocal fold movements, hence the physical intensity of the sounds.

During the production of a glottal stop the vocal folds are shortened and strongly tightened. They are softly tightened and vibrate during phonation. They are moderately forced apart for the production of voiceless sounds and forced even wider apart in aspirated sounds, at least in general.

Figure 4.5 illustrates a simplified diagram of the glottis, the airflow volume and the shape of the spectrum of the resulting buzzing.

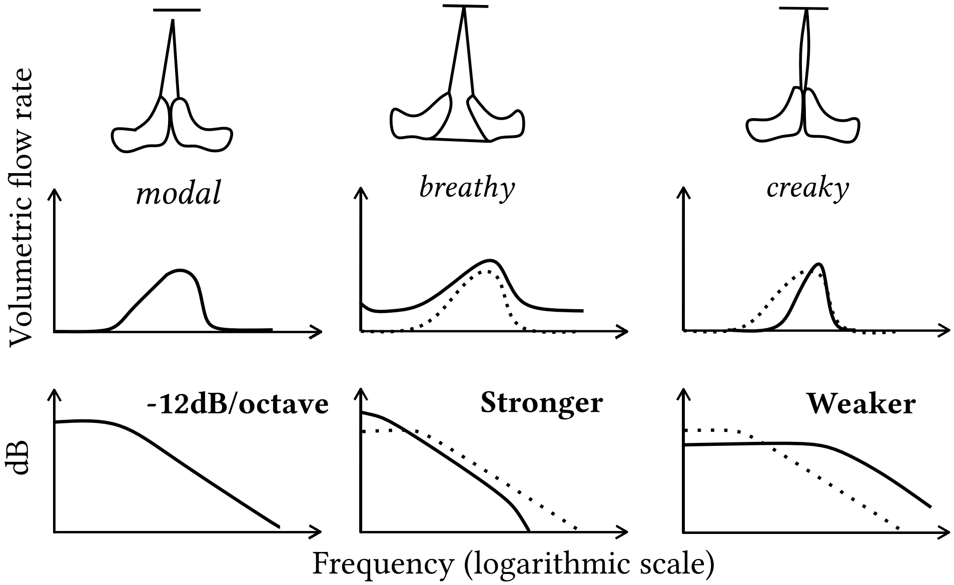


Figure 4.5: *Modal*, *breathy* and *creaky* voices (adapted from a figure by K. N. Stevens): vocal folds (top), airflow (middle) and the spectral slope of the sound sources (bottom)

There are several types of phonation. During *modal* phonation, the opening time of the glottis is slower than the closure time. The spectral slope (see Chapter 5) is minus 12 dB per octave at the source. In contrast, for a *breathy* voice, the closing gesture of the glottis is often incomplete, and less rapid than in the modal voice, so that the vocal fold movement is more symmetrical. The acoustic consequence is a stronger spectral slope, which means that the higher harmonics have lower amplitude. The *breathy* voice is moderately intelligible. In producing a creaky voice, the arytenoids are drawn together, and the vocal folds can vibrate only over part of their length. The medium and high frequencies are strengthened. The glottal closing gesture is particularly effective, and the creaky voice is usually well timbered.

In the world's languages, there are two ways of using of the different states of the glottis: either to contrast phonemes or to express gender, emotions or attitudes. Firstly, the different states can be used to differentiate modal, creaky or breathy vowels and consonants (i.e. phonological function). In voice-register languages such as Mon (Môn-Khmer family), two words can differ because of their different vibration modes of the vocal folds despite their identical phoneme sequence. That is, one of them is produced with a modal voice-register (normal, *unmarked*), and the other with a breathy register. Secondly, in many languages, different voice qualities have no phonologically distinctive function. The intentionally *breathy* voice of a woman can evoke seduction or intimacy (as in the singer Jane Birkin's voice) or pathology. Breathiness contributes to the perception of femininity. A creaky voice can evoke anger. Both breathiness and creakiness may also be linked to vocal fold disorders .

3. The third phase is articulation.

   This phase transforms the buzzing created by the vibrations of the vocal folds into speech. The buzzing is composed of the fundamental frequency of vibration and a series of higher frequencies called upper harmonics, which are multiples of the fundamental frequency. The buzzing is filtered in the supraglottal cavities, basically the pharyngo-laryngeal, buccal and nasal cavities. The cavities act as resonator cavities (or resonators). The resonance qualities of these cavities can be modified by the involvement of the mandible, the tongue, the lips, the soft palate, the advancement or retraction of the pharyngeal zone, and the larynx height. The harmonics of the buzzing, the frequencies of which are consistent with the natural resonances of the resonators (or *poles* in the transfer function, which define the relationship between the vocal tract output and input), are amplified (formants) and the others damped (see Chapter 4).

Speech corresponds to an alternating movement of the lowering-raising of the *mandible* and the *tongue* (Ferdinand de Saussure), occurring every 120 ms on average De Saussure (1959).

The jaw lowering movement has to do mainly with the pronunciation of a vowel, and the raising movement with that of a consonant. The degree of constriction from strongest to weakest makes it possible to differentiate 7 types of phonemes: stops (involving complete closure), fricatives and semi-consonants, as well as high, mid and low vowels (involving the least closure). The amplitude of the oscillation movement of each jaw is also controlled by prosody. A greater lowering of the mandible allows a greater precision of the tongue's movements, on the one hand, and an increase in the frequency of the first formant, and thus of the physical intensity, on the other hand. An overly lowered jaw often accompanies the pronunciation of accented vowels, even of high vowels like /i/.

Further readings on speech production: MacNeilage (1983), Harrington & Tabain (2006), Marchal (2011), Redford (2015).

More specifically on the vowels and consonants: Maddieson (1984), Ladefoged & Maddieson (1996), Ladefoged (2005), Rogers (2014),

On the larynx: Vaissière (1994), Kreiman & Sidtis (2013), Esling et al. (2019).

5 The speech signal and acoustic phonetics

Acoustic phonetics deals with the physical properties of the speech signal transmitted from the speaker's mouth to the listener's eardrum. In this chapter, we will present some information on the sound wave in general, and then on the sound wave corresponding to the speech signal, which contrary to the sounds of nature, is a product of the human vocal tract.

The physicist and physiologist Hermann von Helmholtz, Von Helmholtz (1867) established the scientific basis for the analysis of the acoustic signal and its perception. At the end of the 19th century the Fourier transform of Joseph Fourier, a mathematical function discovered by the Baron of the same name, permitted the decomposition of any complex wave into a series of sinusoidal elementary waves with different frequencies, amplitudes and phases. The invention of the *telephone* and the *microphone* by Graham Bell in 1876-1878, the *modern tape recorder* by Pfleumer (1928) and the *spectrograph* in 1941, during the Second World War, followed by the development of *voice technologies* in the 1960s (formant synthesis in 1960, speech recognition as early as 1952 and signal processing on a computer marked the entry of the *acoustic* dimension into phonetic studies, (see a review by Holmes & Holmes (2002)

The discovery and the quantitative description of the acoustic effects of coarticulation phenomena between phonemes in sequence began in the 1950s. In 1952 an article by Gordon Peterson and Harold Barney on English vowels masterfully illustrated the relationship between the perceived timber of the vowels and the value of their first three formants, as well as the acoustic variability of the vocal productions of men, women and children, Peterson & Barney (1952). At the same time, the book *Preliminaries to Speech Analysis* by the linguist Roman Jakobson, the Swedish speech acoustician Gunnar Fant and Morris Halle, Jakobson et al. (1952) examined the acoustic correlates of the distinctive features, of which the very small inventory would make it possible to characterize all the distinctive differences used by the world's languages. The book constitutes a very important land-mark in the study of the concrete, auditory and perceptual correlates of the phonological, abstract, distinctive features and in the relationship between phonetics and phonology. In 1960, Gunnar Fant's *Acoustic Theory of Speech Production*, in line with the earlier work of the Japanese phoneticians by Chiba & Kajiyama (1941) explained in great detail the relationship between the shape of the vocal tract and its resonance properties on the basis of data coming from x-rays. It is to Kenzo Ishizaka and James Flanagan, Ishizaka & Flanagan (1972) that we owe the first model of vocal folds. As early as the 1970s, Ken Stevens's work at MIT and his search for invariant acoustic correlates of phonetic features fueled controversy over the existence of absolute acoustic invariance in the realization of features (namely *Stevens's theory of invariance*) in spite of the great variability observed in the realization of each phonemic feature, Stevens & Blumstein (1981). Stevens proposed a second important theory: the *Quantal Theory*. According to Stevens's *quantal Theory*, a phoneme would be all the more common in the world's languages if only its acoustico-articulatory properties were stable and not sensitive to small articulatory variations in its realizations, Stevens (1989).

Waves are propagations of pressure changes produced by the vibrations of particles in the surrounding environment. The environment is atmospheric air for humans and water for fish. When the air particles are at rest, they are equidistant and move rapidly in all directions. A shock sets them in motion, creating alternations of zones of local air rarefaction and suppression. The propagation of the pressure changes is rapid, about 340 m/sec at a temperature of 20°C. The pressure changes are transformed into mechanical vibrations in the eardrum (see Chapter 8).

The acoustic properties of sounds that underlie phoneme contrasts include (see Stevens and Fant for a review, Fant (1960a), Stevens (2000) are

- Their *acoustic duration* of the events and silent interval (related in part to the *perceived length* and different cues for the identification of the phonemes);

- Their *fundamental frequency* (abbreviated F_0 which is detectable if the sounds are periodic or quasi-periodic (i.e. if the sound is voiced); F_0 is related to the *perceived pitch*; the *perceived voice quality*, such as breathy or creaky, depends on the exact *vocal fold vibration pattern*, and to the variations that occur in the fundamental frequency, i.e. perturbation of fundamental frequency from cycle to cycle (*jitter*), and disturbance of amplitude of sound wave (*shimmer*); see a review, Kreiman & Gerratt (2003);

- Their *physical intensity*, related to the *perceived loudness* for the vowels. *perceived loudness* depends mainly on the overall acoustic amplitude, but also to the *spectral tilt* related to the relative distribution of energy in low, medium and high frequencies (on the effect of the spectral balance on the perception of lexical stress, Sluijter & Van Heuven (1996);

- Their *spectral composition*, and formant-frequency pattern related to the *perceived timber* of the vowels, in relation to the dominant frequencies in the energy distribution in the frequency scale (the formant frequencies for vowels and energy distribution in the frequency scale for noises) (see Chapter 8;
- The *formant transition* in the vocalic portion of the syllables, as a cue for the identification of the place of constriction of the consonants
- Their *relative amplitude of the formants*, related to the distinction between oral and nasal vowels (Delattre (1966b);
- The *rise-time of the fricative noise* to differentiate affricates (abrupt attack of the friction noise) from fricatives (soft attack of the friction noise);
- The *noise intensity*, as a secondary cue for distinguishes strident (/s/ and /ʃ/) and non strident fricatives (/f/ or /θ/),
- The *presence or absence of a release burst* as a cue to distinguish between stops and non stops consonants;
- The *stationary* or *dynamic aspect* of the resonances (vowel diphthongs, vowel diphthongisation, transitions from vowel to consonant as a cue to the identification of the place of articulation of the consonants, etc.);
- Their *rise time of the first formant* at voice onset for the vowel, after a silence, (there is no shift if a vowel follows the silence , and rapid elevation of the first formant if a consonant follows the silence;

The particular feature of *speech* sounds as compared to other sounds is that they are the product of a source (essentially the *human vocal vocal folds*) and the filtering of that source by the *human vocal tract* and are interpreted as such by listeners. Figure 5.1 illustrates the *source-filter theory*.

- 1) *Spectrum source*: All voiced sounds have as their origin glottal buzzing, produced by the vibrations of the vocal folds. The spectral slope at the source is approximately -12 dB at the source (for a modal phonation, the phonation type by default, see Fig. 4.5).The spectral slope is steeper in the case of a breathy voice and more gradual in the case of a creaky voice. As

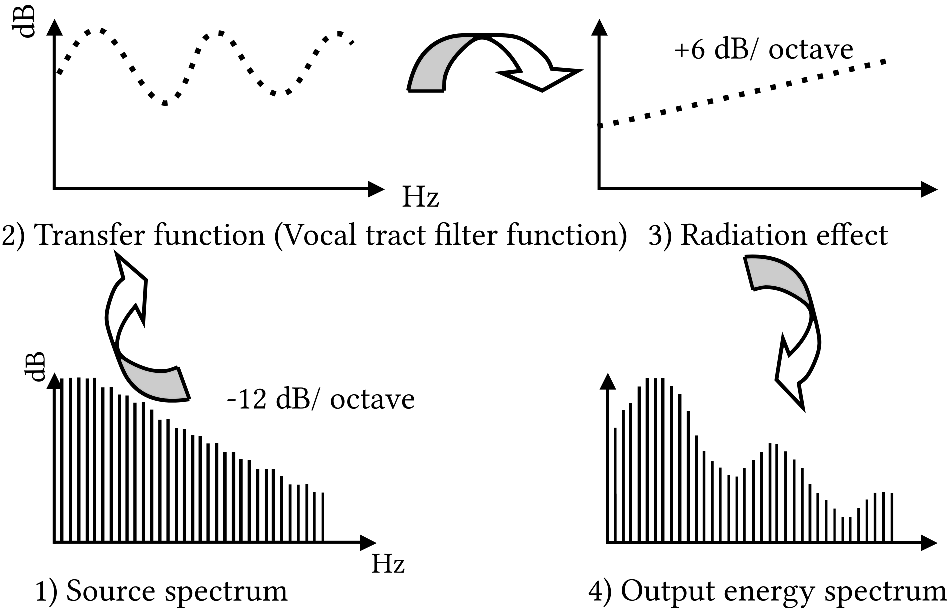

dB
Hz
+6 dB/ octave
2) Transfer function (Vocal tract filter function)
3) Radiation effect
dB
-12 dB/ octave
Hz
1) Source spectrum
4) Output energy spectrum

Figure 5.1: The schematic representation of the source-filter theory of speech production. (1) The source spectrum created by the vibrations of the vocal folds; (2) the voice source is filtered through the vocal tract. (3) the radiation component at the lips rises the global spectral slope; (4) the resulting spectrum.

mentioned before, the buzzing (source signal) is composed of the fundamental frequency, which corresponds to the frequency of vibrations of the vocal folds, and the frequency values of the harmonics, which are the integral multiples of the fundamental frequency. For example, if the vocal folds vibrate at a rate of 120 times per second (the mean value for an adult male), the source signal will consist of the following harmonic frequencies: 120 Hz, 240 Hz, 360 Hz, 480 Hz, 600 Hz, 720 Hz, 840 Hz, etc.

- 2) *Transfer function*: This buzzing excites the vocal tract (1 in Figure 5.1), which has an enclosed volume. Every enclosed volume has natural resonances (represented by its transfer function) and the enclosed volume corresponding to the vocal tract is modifiable by the movements of the articulatory organs. There are about four natural resonances below 4000 Hz (4500 Hz in a woman's vocal tract, which is shorter than that of a man, mainly because her larynx is higher, and a shorter vocal tract produces higher resonances). A vocal tract in neutral position, without constriction, as for the neutral vowel /ə/, can be modeled by a uniform tube closed at one end (representing the glottis) and open at the other (the lips). If this tube has a length of 17.5 cm (which corresponds approximately to that of a male vocal tract), the natural resonances will be 500 Hz, 1500 Hz, 2500 Hz and 3500 Hz. When filtering, the harmonic zones corresponding to the natural resonances of the vocal tract are reinforced and the others are attenuated (2 in Figure 5.1). A reinforced harmonic zone, where the energy is concentrated, is called a *formant*. We refer to the formants of a given vowel by numbering them from the one with the lowest frequency: F_1, F_2 F_3, etc. For the neutral vowel, F_1, F_2 and F_3 are approximately 500 Hz, 1500 Hz and 2500 Hz, respectively. Formant frequencies indirectly inform the listener of the shape of the cavities that have made them.

- 3) The spectral slope of the resulting signal is raised by 6 dB by the so-called phenomenon of *lip radiation* (3 in Figure 5.1).
- 4) The *resulting signal* is thus the product of the source signal, the transfer function and the radiation effect (4 in Figure 5.1).

There are two types of noise sources.

- Firstly, a *glottal noise source* created at the level of the glottis, which is the only source for glottal aspiration noise in aspirated stops and /h/; the glottal noise source will be filtered by the entire vocal tract; the fact that the glottis is opened will result in a weakening of the energy in the region of the first formant;
- Secondly, a *supraglottal noise source* created along the supraglottal cavity: instantaneous release burst following stop consonant release, continuous supraglottal friction noise created at the level of a strong constriction in fricatives and stops; the supraglottal noise source will be filtered by the cavity in front of the constriction.

The periodic buzzing produced by the vibrations of the vocal folds (voice source) is the main source for most of the sounds, since most of the sound are voiced. Voiced sounds are much more audible than whispered speech.

Formant frequencies depend, among other things, on the length of the cavities (and to a lesser extend on their configuration). When a bottle of water is filled, the noise produced by the squirting of the water becomes more acute as the bottle fills up: the smaller the space occupied by the air, the higher its natural resonances. A very acute sound from the bottle tells us that it is time to close the faucet! Similarly, the smaller the vocal tract, the higher its natural resonances. A vocal tract twice as short (as in a child compared to an adult man) has natural resonances twice as high. For a vocal tract of 8.75 cm, F_1, F_2 and F_3 are approximately 1000 Hz, 3000 Hz and 5000 Hz, respectively.

In the same way, the shorter the cavity in front of the constriction (that is, the more the constriction place is anterior), the more the noise of the fricatives will be of high frequency (the excited resonances are basically those of the cavity situated in front of the constriction, as mentioned before). For example, the noise is more acute in the realization of /s/ than in /ʃ/; for the latter, the protrusion of the lips and/or the withdrawing of the tongue allow the lengthening of the front cavity and thus lowers the noise level. The noise of back fricatives is grave: the more posterior the constriction, the darker the timber of the fricative.

The modification possibilities of the resonances are anatomically constrained. The lowest resonance, namely F_1, may vary for a male speaker between 150 Hz (in the case of total closure of the vocal tract for stops), 200 Hz (for a high vowel) and 800-1000 Hz (for the lowest vowel); F_2 between 750 and 2500 Hz, and F_3 between 1500 and 3400 Hz.

Formants are all modified by the general shape of the vocal tract but some of them are more sensitive than others to the movements of certain articulators, depending on the global configuration of the vocal tract.

Three acoustic principles exert a decisive influence on spectral characteristics.

- Firstly, the *frequency* of each formant cannot be controlled in a strictly independent way. Generally speaking, the longer the back cavity is, the shorter the front cavity. That is, all other things being equal, a decrease in the frequency value of F_1 results in a decrease in the frequency value of F_2 for back vowels. The proximity of F_3 and F_4 (for example in the French /i/), which allows the production of a large amount of energy (spectral prominence) at about 3000-3200 Hz, is possible only if F_1 is very low. The more open the vocal tract, the more difficult it is to round the lips, and therefore to manipulate the formants mainly due to the front cavity.

- Secondly, *physical intensity* is mainly due to the contribution of the frequency of F_1: all things being equal, the low vowel /a/ will be the most intense vowel because it has the highest F_1 frequency value, and the high vowels /i/ and /u/ with very low F_1 frequency value are not very intense vowels and are more prone to be transformed or to disappear (as evidenced by sound changes): the final ɑ of Latin did not disappear but was transformed into the silent e in French (*bonna* [bona] > bonne [bɔnə] "good") and all the other vowels in word-final position completely disappeared (*bonus [bonus]* > *bonu* > bon [bɔ] “good”). For a short review, Vaissière (1996).

- Thirdly, the perceptual salience of a formant can be modified, and this depends on the proximity of that formant with the surrounding formants. When two resonances (i.e. two formants) are close to each other (which is possible only in the case of a very strong constriction or when the front and back cavities have a very different section size, Fant (1960a), their amplitudes reinforce each other, a dominant spectral peak is created and as a result their auditory salience is reinforced. The French canonical /i/ is characterized by a mutual reinforcement of the amplitude of F_3 and F_4 towards 3000-3200 Hz for male speakers (F_2 does not reach a perceptible level when F_3 and F_4 are grouped), while for the French /y/ we can observe a mutual reinforcement of the amplitude of F_2 and F_3 towards 1900-2000 Hz (here F_3 does not reach a perceptible level). The vowels with two formants close to each other, and which are thus reinforced, are called *focal vowels*, Stevens (2000). Conversely, the connection of the oral cavity to a lateral cavity (for example during a nasalization process) or to the subglottal cavity (as in the case of a breathy voice) will cause anti-resonant frequencies (and additional resonances) and will consequently reduce the amplitude of certain formants or shift them in the frequency scale. More efficient closure of the vocal folds increases the intensity of the higher harmonics. All of the speech organs, including phonatory and articulatory organs, contribute to increasing the acoustic contrast between certain phonemes in the language. The phonetic realization of the system of opposition between phonemes in a language is often simpler to describe phonetically in terms of acoustic differences than from an articulatory perspective in terms of differences in the configuration of the articulators, which are more difficult to quantify, and compensatory gestures between the articulators, such as the tongue and the jaw exist (See chapter 7). For example, the elevation of the larynx, which is difficult to measure, results in a decrease in the length of the back cavity and thus an increase in the frequency of the mid-wave resonances, assoicated to the back cavity (as in the F_2 of the French /i/), and the effect on the formant frequencies is easy to measure,

Traditional articulatory descriptions that take into account only the tongue and the lips, or the acoustic descriptions of vowels through triangular or quadrilateral vowel diagram that rely only on the first two formants, for example, are not sufficient either for basic or for applied research. Learning how to decode spectrograms is very useful for speech researchers and it is pleasant way to start phonetics for the novice.

Spectrograms make speech visible. A spectrogram is a three-dimensional visual representation of sounds that allows the examination of the basic acoustic properties of sounds. Figure 5.2 illustrates the spectrogram corresponding to the phrase *Voici une poignée de noisettes …* (Here is a handful of walnuts and hazelnuts) pronounced by a French male speaker. The horizontal axis represents time (each mark representing 100 ms), and the vertical axis represents frequency, here marked from 0 to 7000 Hz in Figure 5.2. To display the frequencies up to 4000-5000 Hz is sufficient to study of the acoustic characteristics of the vowels. The visibility of higher frequencies is necessary for studying the acoustic characteristics of the release burst, and the friction noise of the stops and fricatives (Look at the realisation of /s/ and /z/ in Figure 5.2). The shading of the plot shows the energy distribution in the frequency scale, in relation to the intensity of the spectral components and therefore the formants (and the main resonance of noises). There is no energy at all (a silence) during the closure of the consonant /p/ and /t/ and note the presence of energy in the very low frequencies for /d/. The figure on the top represents the fundamental frequency: it is detected for all sounds, except the unvoiced sounds /p/ /t/ and /s/ as expected. /z/ has been devoiced and there is no detection of fundamental frequency and no voice bar. It is important to note that the phoneme is an abstract notion that does not, strictly speaking, have a physical duration. Nevertheless, the realization of each phoneme, if clearly audible, leaves clear acoustic traces, and the acoustic variability between the different realizations of the same phoneme (which are influenced by a large number of factors, listed above) is easily interpretable by a trained spectrogram reader. For example, the realization of the lip rounding feature corresponding to the French vowel /y/ starts from the first consonant /s/ in the French word *structure* (unlike *stricture*), as discussed previously. In this regard, the acoustic trace of the lip-rounded consonant [s^w] is spectrally explained in terms of the lower resonances of the segment in question due to anticipatory lip rounding for the round vowel /y/, as opposed to the unrounded [s] in *stricture.* A wide-band spectrogram (as in Figs. 8 and 9) allowing the visualization of the harmonic series is of great help if difficulty is encountered in detecting the F_0 in breathy voices or pathological voices.

Figure 5.3 illustrates a spectrographic representation of a few French consonants (standard French, male speaker).

The analysis of a speech spectrogram makes it possible to identify several

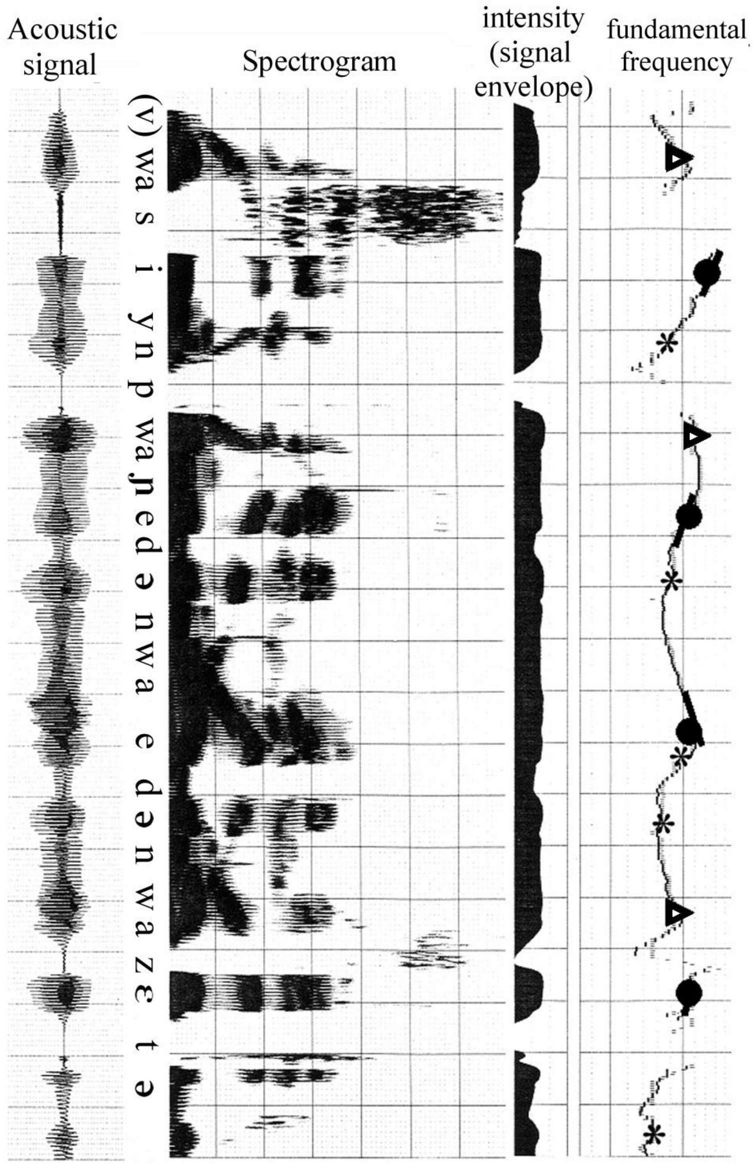


Figure 5.2: Fundamental frequency curve, intensity envelope, spectrogram, phonological transcription and speech signal at the beginning of the sentence *Voici une poignée de noix et de noisettes* … "Here is a handful of walnuts and hazelnuts … "

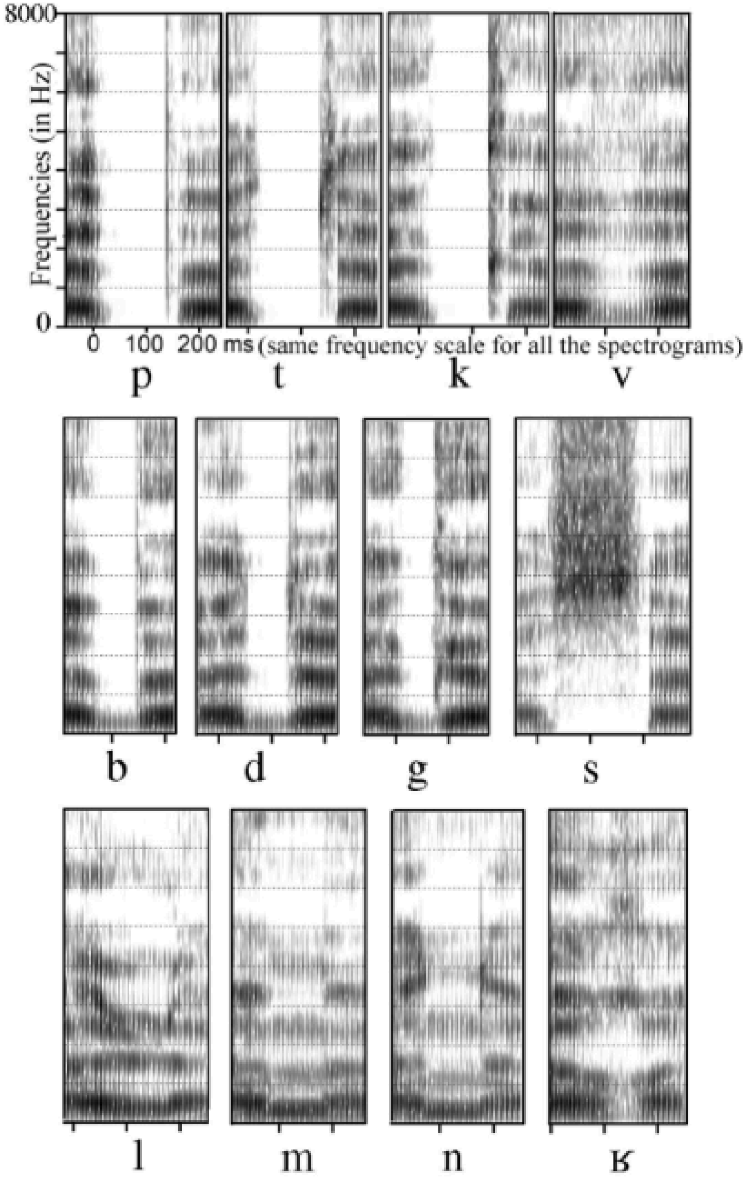


Figure 5.3: Spectrograms of 12 French consonants, each of which is preceded and followed by the vowel [œ]. Fundamental frequency range: 0-10000 Hz.

types of sounds. We invite the reader to identify the acoustic observations described below in the spectrograms.

- Voiced sounds, /b/, /d/, /g/,/v/, /z/, /ʒ/, /l/, /m/, /n/ (and the vowels) are characterized, among other things, by the presence of a *voice bar* in the very low frequencies visible on the spectrogram, by an automatic detection of the fundamental frequency (at the top, in Figure 5.2), and by the presence of quasi-periodic vibrations on the signal.

- Voiceless obstruents /p/, /t/, /k/, /f/, /s/, /ʃ/ are in this way distinguished from their voiced counterparts /b/, /d/ /g/,/v/, /z/, /ʒ/; the partial or total absence of a voice bar in the spectrogram allows the identification of contextually devoiced sounds such as the realization of /b/ in the French words *ro**b**e sale* /ʁɔb sal/ "dirty dress", perceived as *ro**p**'sal* or *ba**g**ue perdue* /bag pɛʁdy/ "lost ring", perceived as *ba**c** perdu* /bak pɛʁdy/ "lost tray". The small valleys on the F_0 curve during the pronunciation of voiced stops and fricatives, called *micro-melody*, correspond to an uncontrolled and transient decrease in the rhythm of the vocal folds' vibration. Micro-melody is due to a pressure increase in the oral cavity (as a result of a supraglottal constriction) and a resulting reduction in transglottal pressure.

- Vowels are identified by the presence of formants in the low and medium frequencies, (4 formants under 4 kHz are expected for a male speaker, for oral vowels). The F_1 frequency value of a vowel (and therefore its energy) increases to a maximum local value on the time axis during its production, and then decreases again toward the next consonant. Vowels have a voice bar, except of course in the case of contextual devoicing: high vowels surrounded by voiceless consonants may be devoiced (as sometimes observed in the realization of the French word *schism* [ʃism] and devoicing of the high vowels /i/ and /u/ between two voiceless consonants is current in Japanese).

- Canonical stops (/p, t, k/ and /b, d, g/) are identified by the absence of energy in the medium and high frequencies during their closure phase, and by an explosion bar at their release. During their closure phase, unvoiced stops generally correspond to silence on the spectrogram, and voiced stop to a voice bar, as metioned before. In French, /p, t, k/ are not aspirated, and as seen in Figure 5.3, there is no aspiration noise. The voice bar for /b, d, g/ is clearly visible.

- Canonical fricatives (/f, s, ʃ/ and /v, z, ʒ/) are characterized by (1) the presence of a continuous noise during their production, made at the level of the supraglottal constriction and filtered in the cavity located in front of the constriction, and (2) the absence of explosion noise at their release. As shown by the spectrograms, the noise of /s/ is intrinsically more intense than that of /z/ since parts of the vocal folds vibrate for voiced fricatives to the detriment of the physical intensity of the supraglottal friction noise, whereas they are entirely open for unvoiced fricatives. Friction noise is absent for /v/ in Figure 5.3: the labial voiced fricative is often realized as an approximant in French.

- Canonical sonorants are characterized by the presence of formants (as in vowels), but they are of lower amplitude than vowels (with no local peak of energy), and nasal and lateral consonants have anti-formants.

Changes in the shape of the vocal tract are progressive and the movements of the articulators are continuous. This means that there is a progressive change in the underlying resonances. There is a strong similarities between of the F-pattern among vowels, semi-consonants, consonants sharing a similar place of articulation , but different in constriction degree an shape (/i/ and /j/; /y/ and /ɥ, /u/ and /w/, /ɑ/ and /ʁ/, /ɜ̆/ and /r/), see Fant (1960a). Vaissière (2007).

Despite the continuous course of the F-pattern, the spectrogram reveals certain acoustic discontinuities. These discontinuities are created either

(1) by the abrupt halt of the *excitation* of certain resonances: the realization of an occlusion or a strong narrowing at a point of the vocal tract causes the sudden creation of a friction noise and the resonances due to the cavity behind the constriction no longer being excited: during the consonant /s/ (see Figure 5.2, the formants higher than F4 are excited, and during the following vowel /i/, the formants lower than F4 are excited.

(2) by the sudden production of anti-resonances — mainly by the addition of a second cavity, for example by a tracheal (subglottic), lateral or nasal cavity,

(3) by the interruption of the vocal folds' vibration (the voice source is suddenly suppressed), or

(4) by the sudden appearance of a *supraglottal noise source* due to a significant narrowing.

On the other hand, the undisturbed continuity of the F-pattern (Gunnar Fant's term) makes the segmentation difficult between two successive vowels (hiatus vowels, i.e. not separated by a consonant) or between a vowel and the following coda consonant, or between a vowel and a semi-vowel, or between different types of noises, such as the friction noise of /t/ and the devoicing at the beginning of /i/ in [ti], or the friction noise, the aspiration and the devoicing at the beginning of [i] in [t^{h}i]. A sequence such as [nwa] (see Figure 5.2) is difficult to segment.

Formant transitions (particularly F_2 and F_3) at the beginning of the vowel convey information on the place of articulation of the preceding consonant and on the form and position of the tongue during its realization. Bilabials and non-palatalized labiodentals are characterized by their low resonances, that is, they have lower frequency values than those of the following vowel, and consequently transitions from consonant to vowel rise. Non-velarized dental and alveolar consonants are characterized by an F_2 around 1600-1800 Hz (the place of constriction is relatively fixed). The F_2 transitions from a consonant into a vowel therefore descend if the F_2 of the vowel is lower than 1600-1800 Hz, otherwise they rise. The place of articulation of the velar consonant /k/ depends on the vowel that follows, which means that it is phonetically realized as velarized in /ku/, palato-velarized in /ka/ and palatalized in /ki/ (See Chapter 7). In Chapter 8 on ‘perception’ we will see that the modifications of the velar place of articulation as a function of the following vowel are undoubtedly largely due to perceptual constraints and not exclusively articulatory cases, such as the ease of articulation, which is often evoked in the literature. That is, the noise level at the moment of release should be in a precise relationship with the F_2’ (see Chapter 8 on perception for the notion of F_2’) of the next vowel so that /k/ or /g/ is perceived. It is therefore necessary to adjust the length of the cavity in front of the constriction point during the”velar” consonant.

The place of constriction of the consonant does not by itself definitely determine the formant values at the beginning of the following vowel. In this regard, the shape of the tongue plays a key role. If the consonant /k/ or /g/ is palatalized (the tongue moves forward, as for the vowel /i/). The onset of F_2 of the consonant does not vary and it is around 2000 Hz, regardless of the place of articulation of the palatalized consonant (labial, alveolar or velar) or its manner of articulation (stop or fricative). Therefore, for all of the palatalized consonants, such as [p’], [t’] and [k’], [f’], [s’] or [ʃ’], the F_2, the formant transitions towards the vowel are almost identical and begin at 2000 Hz. The shape of the formant transition does not allow to recover the place of constriction of the consonants. The F_2 transitions are ineffective as clues to distinguish the different places of articulation. [p’], [t’] and [k’] can be differentiated perceptually only from the height of the friction noise at the release, and [f’], [s’] and [ʃ’] from the height of the friction noise during the constriction of the vocal tract. Palatalization is phonemic in Russian but is often contextual in most languages, when the consonant is followed by the vowel /i/. The acoustic distance between /ti/, /ki/, /si/, /ʃi/ is reduced Compare the pronunciation of the Latin word ”natio”, with the French prononciation /nasjɔ̃/ ad English ˈneɪʃn̩ /°.

Thorough training in acoustic phonetics can be acquired without prior knowledge in physics. A computer, transportable in a classroom or the field, and easy access to software analysis programs, such as Praat, Boersma (2011) , and software synthesis programs (such as Denis Klatt's formative synthesis program, Klatt (1980) or Shinji Maeda's articulatory synthesis program Maeda (1982), Birkholz et al. (2006), downloadable for free on the Internet, can facilitate thorough training in acoustic phonetics and good intuitive understanding of the relationship between articulatory, acoustic and perceptual properties of the signal. See also the application *Cleanaccent* on the web, created by the present author to learn how to read spectrograms and use this ability in different ways.

Readings in acoustic phonetics: for beginners on acoustic phonetics: Fry (1976), Vaissière (2007). Ladefoged (1996).

intermediate level: Lehiste (1967), Fant (1960a). Kent & Read (1992), Pickett (1998), Johnson (2012),

advanced level: Fant (2006).

Stevens (2000).

6 Vowels

The number of vowels in languages varies from one to more than twenty. Most languages have five to seven. French has a different number of vowels depending on the region where it is spoken. More than 99% of languages have at least two vowels. The most frequent phoneme inventory in the world's languages has five vowels (22% of the languages of the UPSID database, Maddieson (1981). 80% of languages have three to ten vowels. The most frequent vowels are, in descending order, the following: /a/, /i/, /u/, /e/, and /o/. Languages tend to exploit only two dimensions of the aperture of the vocal tract (mainly related to F_1) and the degree of frontness/backness of the tongue (mainly related to F_2) for the first eight vowels. The back vowels are generally rounded.

Vowels have a dual nature: articulatory and acoustic and a perceptual dimension.

From the *articulatory* point of view (Figure 6.1), the tongue is the main organ of articulation of vowels. It is raised towards the front of the oral cavity for front vowels (called also palatal or clear vowels) (/i, e, ɛ, a/) and moved backwards for back vowels (also called velar or dark vowels: /u, o, ɔ, ɑ/). The distance between the tongue's surface and the palate increases in the passage from /i/ to /a/ (front vowels) and the constriction shifts from the velar region to the pharyngeal area in passing from /u/ to /ɑ/ (back vowels), which involves passing through /o/ and /ɔ/.

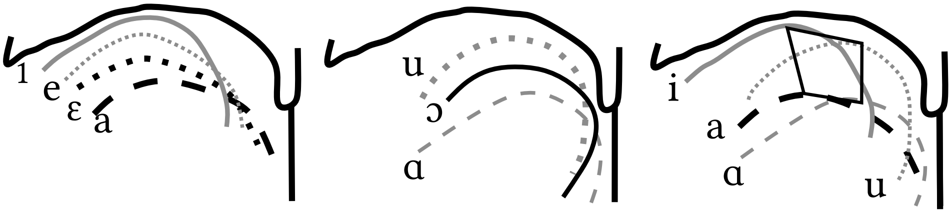


Figure 6.1: Articulatory positions of the tongue for French vowels (according to Straka's X-rays, Straka (1965) and the corresponding vocalic trapezoid

From the *acoustic* point of view (Figure 6.2), the configuration of the *lips* plays an important role in modifying the length and therefore the resonances of the front cavity. F_2 is often a resonance of the front cavity (as in back vowels), but F_3, too, may be associated with the front cavity, as observed in the realization of /i/. For example, the vowels /i/ and /y/ are acoustically distinguished by F_3: F_3 is high for the spread vowel /i/ and it low for the round vowel /y/). In the case of /i/, the front cavity is short, and F_3 gets close to F_4): the regrouping of F_3 and F_4 results in a strong spectral peak in the F3-F4 region (around 3000 Hz for a male speaker). In the case of /y/, the front cavity is longer (longer than the back cavity), and F_3 gets close to F_2): the regrouping of F_2 and F_3 results in a strong spectral peak in the F2-F3 region (around 2000 Hz for a male speaker).

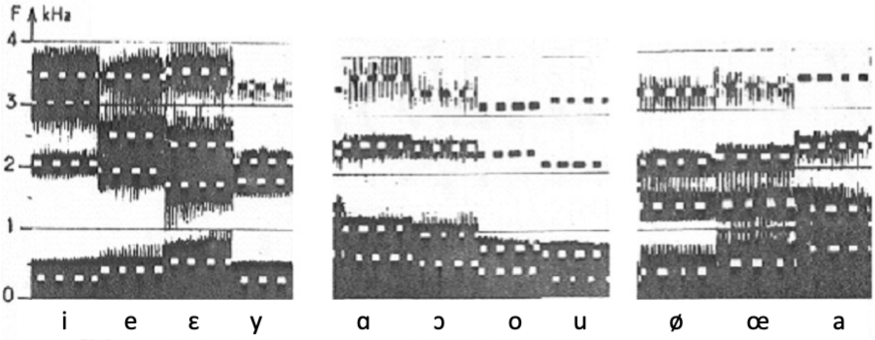


Figure 6.2: Spectrograms of typical oral vowels in French (excerpt from a figure of a book by Jean-Sylvain Lienard,Liénard (1977) on the left, four acute vowels (with F_2 higher than 1800 Hz); Mid: the grave vowels (with F_2 lower than 1200 Hz); right: the vowels neither acute, nor grave (F_2 around 1500 Hz).

The action of the lips makes it possible to distinguish in French between *pie* /pi/ “magpie” and *pu* /py/ “could”, *fée* /fe/ “fairy” and *feu* /fø/ “fire”, *air* /ɛʁ/ “air” and *heure* /œʁ/ “hour”. The lowering of the soft palate allows the creation of a subsystem of nasal vowels (three in contemporary French, those of the words *paon* /pɑ/ “peacock’, *pain* /pɛ/ “bread”, and *pont* /pɔ/ “bridge”).

Some remarks about the acoustic criteria for separating acute and dark vowels and about the correspondance between the acoustico-perceptually characteristics of the vowels and their usual distinctive features. Figure 6.2 illustrates the spectrograms corresponding to the eleven oral vowels of French. Note that Daniel Jones choose the French vowels as representive of the cardinal vowels Ladefoged & Johnson (2014).

1. A large distance between F_1 and F_2 characterizes the acoustico-perceptually *clear or acute* vowels (sometimes also called ”bright”, because of the color their timber evoke) (on the left in Figure 6.2). The F_2 in this type of vowel is high (more than 1800 Hz for a male speaker) and the energy is distributed in the high frequencies (see Figure 6.2, left side) F_3 est strongly excited, and even F_4, except in the case of /y/.

   Formants higher than F2 have a perceptual weight for the acute vowels, Carlson et al. (1970). The articulatory unrounded front vowels /i/, /e/, /ɛ/ or /æ/ are acoustico-perceptually acute (i.e. in our acsoustic definition of acuteness, an acute vowel is realized as acute if F_2 is higher than 1800 Hz).

2. The acoustico-perceptually *dark or grave* vowels (see Figure 6.2, center) are characterized by the proximity of F_1 and F_2 below 1000-1200 Hz (for a male speaker) and by a low perceptual weight of the formants higher than F_2 That is, their F_2 is lower than 1200 Hz for a male speaker. The articulatory rounded back vowels /u/, /o/, /ɔ/ or /ɑ/ are acoustico-perceptually grave (F_2 lower than 1200 Hz). The vowel color of the back vowels can be synthesized using formant synthesis by a single formant corresponding to the grouping of the tow lowest formants Delattre et al. (1952).

3. The acoustico-perceptually *central* vowels in which the energy is evenly distributed (see Figure 6.2, right side) are characterized by their F_2 located around 1500 Hz. The rounded front vowels /œ/ and /ø/ are acoustico-perceptually central (F_2 around 1500 Hz) in terms of their formant pattern, although they are considered as +front, + round in terms of phonological distinctive features (see discussion later).

The vowels displayed in Figure 6.2 are hyperarticulated and have been pronounced with care. In continuous speech, back vowels such as /u/ and /o/ are subject to tongue fronting (their intrinsically low F_2 becomes higher) when they are surrounded by consonants pronounced towards the front of the vocal cavity: /u/ tends toward /y/ and /o/ toward /ø/ (vowel). (For the effects of the consonantal context on the formant values of the vowel, House & Fairbanks (1953). If they are short, they are acoustically centralized and/or more strongly assimilated to their surrounding consonants, depending on their prosodic position (see Chapter 9). The highest F1 value for French /a/ is not obtained when /a/ is pronounced in isolation, but when /a/ is surrounded by the (only) back consonant /ʁ/.

Figure 6.3 shows six configurations of the vocal tract generated by Maeda's articulatory model. Using articulatroy synthesis allows to have a clear, unbiased idea of the acoustic consequences of a change in configuration of the lips and of the tongue, and also to the regions of the vocal tract which are the most sensitive to a slight configuration change (the so-called sensitivity functions, Fant & Pauli (1974).

For six French vowels, the configuration of either the lips or the tongue, has been changed: the lips for the pair /i-y/ (allowing F3 to get close to F4 for /i/ and close to F3 for /y/). A change in the lip configuration is sufficient for creating the pair /œ-ø/ (rounding ot the lips allowsthe lowerig of all formants), no change in tongue configuration is necessary.

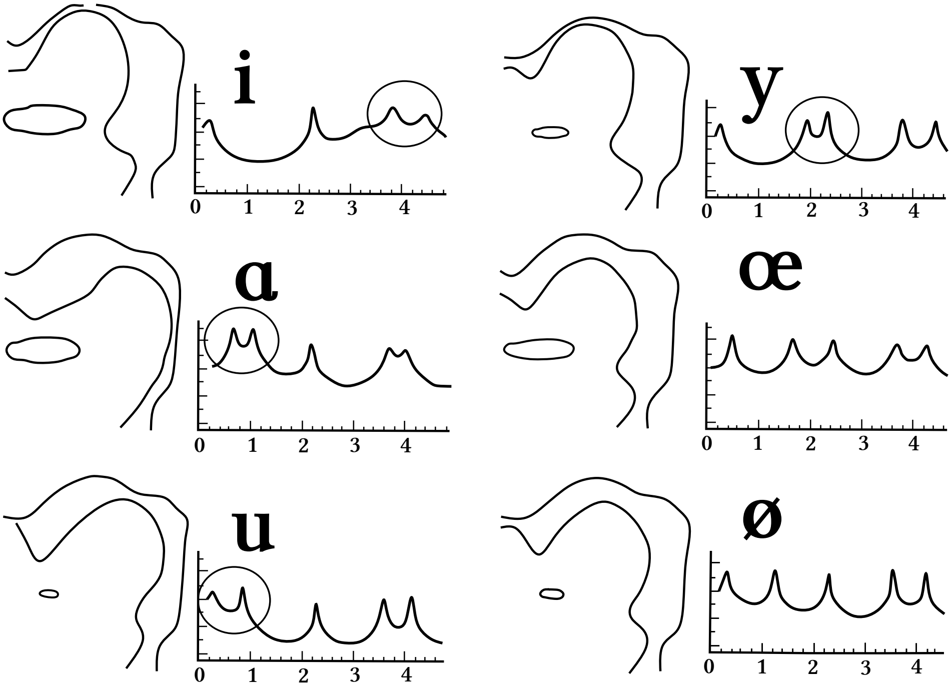


Figure 6.3: The modeling of six vowels through articulatory synthesis. Left: the vocal tract configuration and the lips configuration. Right: the resulting spectrum.

In languages where labiality does not play a distinctive role for vowels, the front vowels are generally not rounded and the back vowels are rounded in more than nine out of ten languages. Why? As mentioned before, lip spreading reduces the length of the front cavity, thus increasing the frequency value of F_2 and/or F_3 and the increase is greater for the formant mainly affiliated to the front cavity; conversely, lip rounding and/or protrusion reduces the formant frequency values (generally F_2 and/or F_3 and the increase is greater for the formant the mainly affiliated to the front cavity (F_2 is mainly associated with the front cavity for back vowels). Lip spreading in front vowels and lip rounding in back vowels increase therefore the perceptual contrast between the two groups of vowels, by increasing the distance between F_1 and F_2. The lips and the tongue collaborate to distinguish acute (front) and grave (back) vowels. The acoustic realisation of the distinctive feature [+/− back] (back/front vowels) is not only a matter of the position of the tongue in the mouth: backing of the tongue is not enough to create the timber of /u/, /o/ and /ɔ/: rounding of the lips is necessary. Furthermore, the amount of rounding is adjusted to the length of the back cavity: the effective length of the front cavity and the effective length of the back cavity have to be similar to create two resonances of about the same frequency value and create a spectral peak (on effective length, Fant (1975).

There are various possibilities of *compensation* between the articulators Maeda (1990). , which are often taken into consideration by clinical phoneticians. For example, the mandible generally (but not necessarily) accompanies the tongue movements to increase openness and to achieve a particular acoustic target, such as an high F_1: the lower the mandible, the lower the tongue, thus producing a higher F_1. The pipe smoker compensates for the immobility of the mandible by more ample movements of the tongue, Gay et al. (1981). The degree and position of the constriction, the lowering of the mandible and the configuration of the lips and the height of the larynx allow a large set of compensatory gestures (or reinforcing gestures). Individual speakers (also speakers of some dialects or some languages) make relatively wider use of the mandible height or the lip configuration than others. The compensatory gestures reinforce the idea of the *primacy of the acoustic goal on articulation* for the identification of phonemes by the listener. (on individual differences in vowel production, Johnson et al. (1993).

This is in line with the notion of language-specific global *articulatory habits* (or *settings*, or *basis of articulator*). *Articulatory setting* is defined as a general habit of speech community, at the level of the supraglottic articulators and at the phonatory level. Eduard Sievers is considered theprecursor of the concept of articulatory basis, Sievers (1893). The notion of *Articulatory setting* was important for the phoneticians of the beginning of the last century in order to explain sound changes Sapir (1925), Gick et al. (2004). For an historical survey of the concept of articulatory settings and extension ot the laryngeal mode, Laver (1978), Schweyer (1987).

Pierre Delattre proposed a number of principles related to the notion of general *Articulatory setting* for English learners of French Delattre (1951), see also Matte (1982). Pierre Delattre proposed three opposite modes for French and English.

1. The first modes corresponds to *muscular tenseness*. Tenseness induces relative stability during articulatory states, no diphtongization of vowels, less affrication of consonants, less weakening of vowels and consonants in weak position, more extreme vowels, etc.). Delattre considered French as tense, and English as lax, leading to unstable articulatory states and diphtongization, etc.

2. The second mode corresponds to *open syllabification*. open syllabification (called also rising mode) leads to gradual vocalic onset, to vocalic anticipation during the consonant preceding the vowel, consonant in coda to detach itself from the preceding vowel and to link with the following syllable (”enchaînement”), clear release of final consonant if not linked to the following syllable, etc. French corresponds to open syllabification and English to close syllabification: abrupt vocalic onset, closer link between the vowel and the following consonant in coda, no clear release of the consonant in coda, no ”enchaînement”, etc.

3. The third mode corresponds to *anterior articulation* (concave shape of the tongue, no retroflexion, weak lip movement). French is characterized by an anterior articulation, and English by a more posterior articulation, as a general setting.

The trapezoidal (/i-u-a-ɑ/) or triangular shape (/i-u-a/) of the *vocalic triangle* is the geometric figure obtained by connecting the highest points of the tongue for each vowel. A distinct trapezoid appears in languages that have two /A/, a grave (back) [ɑ] and an acute (front) [a]/[æ]). whereas a triangle appears for those that possess only one very low vowel of [a] timber. The elegant correspondence between the representation of vowels by their first two formants (called the *vocalic acoustic triangle* or *vowel chart*) and the articulatory vocalic triangle is misleading since it neglects the decisive effect of the lips and the perceptual weights of the formants higher than F_2 on the timber of front vowels. Acoustically, from a theoretical point of view, it is possible to deform the trapezoid by taking as a reference point the maximum constriction point rather than the highest point of the tongue (a notion however difficult to apply to high front vowels). This interpretation is in line with Gunnar Fant's nomograms and Stevens's three dimensional model for the vowel : position of the tongue constriction, the size of the constriction formed by the tongue, and the dimensions in the vicinity of the mouth opening, Stevens & House (1955).

F_3 plays an important role in French as well as in other languages. The movement of the lips (i.e. lip rounding) is sufficient to lower the F_3 of /i/ from 3000 to 2000 Hz (and thus produce an /y/), as mentioned before. In French, /i/ (the proximity of F_3 and F_4) and /y/ (the proximity F_2 and F_3) are two focal vowels, which therefore have a precise acoustic definition.

Vowels with medium aperture (i.e. the so-called mid vowels) have a less precise acoustic target: the timber can evolve between /e/ and /ɛ/, /o/ and /ɔ/, /ø/ and /œ/, the acoustic boundaries are not very precise and the opposition is often neutralized in certain positions, like in French.

Most of the world languages are tonal languages and use tone as a distinctive feature to contrast their vowels. Languages with large vowel inventory, i.e. with more than eight to ten vowels use more than the open/close and front/back dimensions. They use at least three acoustic dimensions such as *labiality*, *nasality* or *length*. One fifth of languages oppose oral and nasal vowels, on the one hand, and long vowels and short vowels, on the other, see Vallée (1994).

F_3 Chinese use *tones*: F_0 height and shape allow to contrast the vowels; French, German and Swedish use *labiality* to contrast front rounded and front unrounded vowels: F_3 plays a large role for assuming the contrast; French use *nasality*: the flattening of the energy in the F_1 region then becomes the principal distinctive acoustic correlate of nasality. The relative amplitude of the formants become essential. German and Japanese use *length*, etc.

The shape of the tongue may play a role in shaping the vowel timber. Mandarin Chinese in Beijing, American English and other languages (including Naxi, a rare language spoken in China) have *rhotic* vowels; for example the vowel of the word bird (/ɚ/ in American English). The defining acoustic characteristic of (/ɚ/ is the lowering of the third formant. F_3 is lower than 2000 Hz. The lowering of the third formant can be best achieved by is a triple constriction, at the level of the lips, at the front and the back cavities of the vocal tract. If /ɚ/ is selected by a few number of languages, it is most likely because /ɚ/ is easy to articulate, but most likely because of its intrinsic acoustic singularity, with an F_3 lowers than the ”normal” F_3 range.

If a language uses variations in length (long and short vowels, geminate consonants), fundamental frequency (tones), and voice quality (*breathy* voice or *creaky* voice, for example) to create opposition between phonemes and words, such variations may be less broadly used for lexical stress marking, boundary marking and other prosodic purposes (see Chapter 9). What a language does not use to create opposition between its vowels (or its consonants) or word lexical stress remains available as a means of marking nuances of meaning, for example (see Chapter 9).

7 Consonants

Unlike the vowels, for which the vocal tract is fairly open, the consonants are formed with a more or less constricted vocal tract. The vocal tract is

- completely closed for the oral stops (such as /p/, /b/); it is also completely close for the nasal stops (such as /m/ and /n), but the nasal tract is open,

- enough constricted for the fricatives (such as /s/ and /z/) to create audible friction,

- not enough constricted to create friction for the sonorants, but too constricted to sound like vowels, such as the liquids like [l] and [r], and semivowels like [j] and [w].

While the consonants and the vowels differ essentially by their degree of constriction, the consonants and the vowels sharing a similar configuration of the vocal tract (such as /t/-/s/-/ts/-/i/ or /ʁ/-/ɑ/) have a very similar F-pattern (see the striking similarities of the F-pattern of the nomograms for both vowels and consonants, Fant (1960a), Vaissière (2007).

Languages have an average of 22 consonants in great variety. The number of consonants vary from 6 to 95 consonants in the UPSID database, Maddieson (1981). The most common are the 7 stops /p, b, t, d, k, g, ʔ/ the 4 fricatives /f, s, ʃ, h/, the 3 nasals /m, n, ɲ/, the 3 approximants /l, j, w/, the 2 affricates /ts, tʃ/ and the apical /r/. All languages have voiced sonorants and unvoiced obstruents in their systems. Most of the word languages do not have clusters (a cluster is a group of consonants without intervening vowels).

There 24 consonants in English

The 3 unvoiced stops: /p/ (pad), /t/ (ten), /k/ (cab),

and their 3 voiced counterparts: /b/ (bad), /d/ (dad), /g/ good,

the 5 unvoiced fricatives /f/ (fan), /θ/ (thing), /s/ (sad), /ʃ/ (ship, /h/ (her),

the 4 voiced fricatives: /v/ (van), /ð/ (this), /z/ (zap), /ʒ/ (beige),

the 2 affricates: /tʃ/ (cheap), /dʒ/ (judge)

the 3 nasals: /m/ (man), /n/ (no), /ŋ/ (ring),

the 2 liquids: /l/ (light), /r/ (right), /w/ (white), /j/ (you),

Not all combinations of phonemes are permissible and they are governed by language-dependant phonotactic rules. For example, English and French have clusters. In English, the longest possible initial cluster is three consonants long, C1C2C3 , such as ”splendid” in English and ”splendide” in French. In French and in English, in a C1C2C3 cluster, and according to a common phonotactic rule C1 is obligatorily a fricative, C2 a stop and C3 liquid. (See Lindblom & Maddieson (1988), on phonetic universals in consonant systems).

The four main criteria for classifying consonants are *place of articulation* (from labial to glottal), *voicing* (voiced/unvoiced), *degree of constriction* (from stops to approximants), *nasality (oral and nasal consonants.*

1) The first criteria is the *place of constriction.*

Figure 7.1 illustrates the names of the *places of articulation of consonants,* from the glottis (glottal) to the lips (labial). The moving organs, the tongue or the lips, move toward the fixed parts of the vocal tract (the hard palate for the palatals, and the soft palate for the velars). *apico-dental* means that the tongue apex approaches or touches the teeth, and *lamino-alveolar* means that a closure or narrowing occurs between the edges of the tongue blade and the alveoli.

Figure 7.2 illustrates the configuration of the tongue and the lips for the production of labial, dental and velar stops.

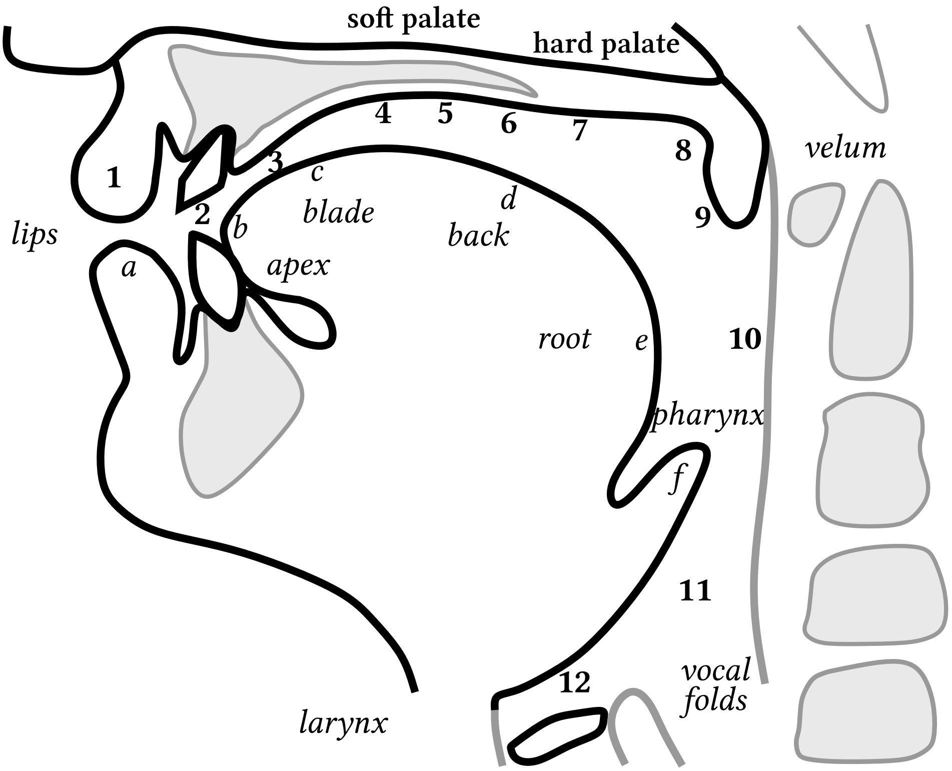
The *shape of the tongue* plays also a role. The acoustic consequences of bunched versus retroflex tongue configuration are not very clear because of compensation phenomena. The different tongue configurations for /r/ produced by speakers of rhotic dialects of North American results in similar first three formant values, Delattre & Freeman (1968). (See Zhou et al. (2008) for discussions).

The *stops* are composed of several phases: a) the closure, b) the instant release of the closure, c) the friction noise (the same for /t/ and /s/) created at the place of constriction, d) the aspiration noise if aspirated, created at the glottis, e) and the formant transition toward the vowel.

What are the cues for the identification of the place of constriction of the stop consonants? The spectral characteristics of the instant release, the height of the following friction noise and its global shape (compact for /k/, diffuse for /p/ and /t/, and the formant transition are cues for the identification of the place of articulation. Ohala et al. (1981).

1. The first cue for the identification of the place of articulation of the consonant is the spectral characteristics of the release and the following friction noise.

   The *”burst”* encompasses the release (instantaneous noise) and about 15 ms of the following friction noise.



1) labial; 2) dental; 3) alveolar; palatal; 4) prepalatal; 5) medio-palatal; 6) postpalatal; velar: 7) pre-velar and 8) velar; 9) uvular; 10) pharyngeal; 11) laryngeal; 12) glottal. a) labial; coronal; b) apical and c) laminal; d) dorsal; e) radical; f) epiglottal

Figure 7.1: The names of the places of articulation of consonants (stylized midsagittal section of the adult human vocal tract)

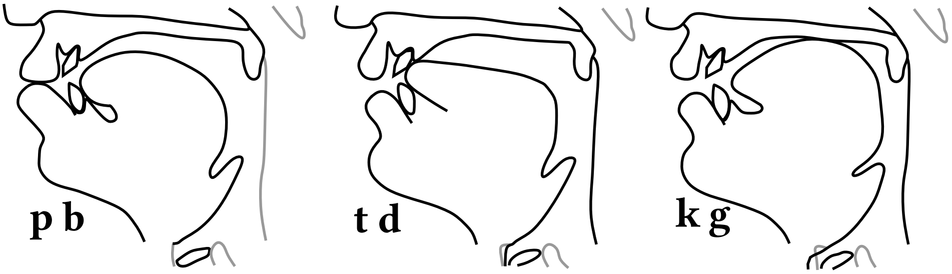


Figure 7.2: Positions of speech organs during the production of oral stops

The spectral shape of the burst for the *labials* is dominated by the energy in the low frequencies, the general shape is diffuse, being falling from low frequencies to high frequencies.

The spectral shape of the burst for the *apicals* is dominated by the energy in the high frequencies, the general shape is diffuse, being rising from low frequencies to high frequencies.

The spectral shape of the burst for the */k/ and /g/* displays a compact burst. Stevens & Blumstein (1978).

The position of the main spectral peak for /k/ and /g/ depends highly on the following vowel:

- grave, around F2 in front of the dark vowels, lower for /u/ than for /ɑ/,

- between F2 and F3 for the not acute, nor dark vowels (/ø/, /œ/, /a/), creating a velar pinch (F2 and F3 are close).

- higher, in front of (F2-F3) for /y/ (i.e. around 2000 Hz),

- even higher, in front of F3 for /e/ (i.e. around 2500 Hz),

- and the highest, acute, near F3-F4 for canonical /i/ (i.e. around 3000 Hz). Fant (1973), Vaissière (2007).

The common *grave* burst in /pu/ and /ku/ leads to their confusion and sound changes (see ohala for examples of sound change based on the acoustic characteristics of the sound Ohala et al. (1981).

The common *acute* burst in /ti/ and /ki/ to reduce their acoustic distance. /t/ burst is however rising and diffuse and the noise extends over the region corresponding to the formants superior to F5 while /k/ burst is more compact and the noise does not extend over the F5 of the vowel. In other words, the acoustic distance between /p/, /t/ and /k/ depends on the following vowel: /p/ and /k/ are acoustically more similar when followed by /u/ than when followed by i/. /t/ and /k/ are acoustically more similar when followed by /i/ than when followed by /u/.

2. The second main cue is the characteristics is the formant transitions. The formant transitions are extensive for /t/ in front of back vowels than in front of front vowels.

   - In /CV/, where C is a labial, the formants after the release tend to rise.

   - In /CV/, where C is a dental or alveolor, the F2 transition starts from about 1800 Hz. F2 transition will be falling from 1800 Hz if the vowel is not acute

(F2 < 1800 Hz) and flat if the vowel has an F2 close to 1800 hz, and rising in the other cases.

- In /CV/, where C is a /k/ or /g/, there is almost no formant movement if the vowel is acute or grave: /k/ and /g/ anticipate the configuration required by the following vowel. If the vowel is neither acute or grave (such as /a/, /œ/, /ø/) F2 and F3 will joint together: forming the velar pinch. F3 plays a large role in that case, for differentiating between /t/ and /k/: F3 is falling in the case of /t/ (parallel to F2), but rising for /k/), Fant (1973).

What are the cues for the identification of the place of constriction of the fricative consonants?

The main cues for identifying the place of articulation of the fricatives are a) the location and shape of energy noise during the consonant and b) the formant transitions (like for the stops).

The resonances of the F-pattern excited by the friction noise are those of the cavity in front of the constriction: the shorter the cavity, the higher the energy concentration. /s/ is much more acute than /ʃ/, and the friction noise of the back fricatives is grave. The formants superior to F4 are excited in the case of /s/. The formants around F3 are excited in the case of /ʃ/. In the case of the back fricatives (such /x/), the F2 of the F-pattern is excited.

The formant transitions from the fricative to the vowel are the same than those for the stops at the same place of constriction /f-v-p-b-m/, /s-z-t-d-n/, etc.

The shape of the tongue may vary for a similar place of articulation. /t/ is rather apico-alveolar in English and lamino-dental in French (lamino-alveolars and apico-dental stops are also possible). Figure 7.3 illustrates the difference in typical tongue configuration for English /t/ and French /t/ (but see Dart (1998) for disparate results).

For *lateral* production, as in light /l/, the tongue produces a central constriction by approaching the palatal arch. The tongue blade is lowered and the air passes through both sides, creating two lateral cavities (hence the presence of antiformants).

Approximants are typically continuous consonants without friction, Maddieson (1981).

2) The second criteria is *voicing*.

For the unvoiced obstruents (stops and fricatives) the glottis stay open and there is no vibration of the vocal folds, and no energy at all. For the voiced obstruents (stops and fricatives) the vocal folds come together and they vibrate if

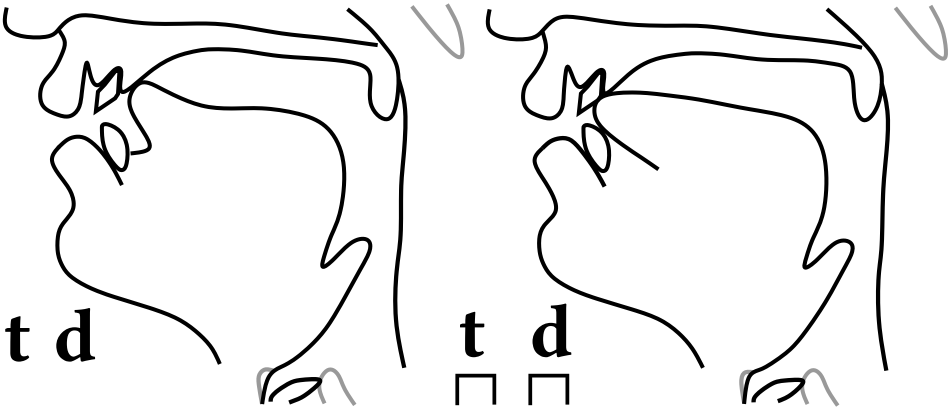


Figure 7.3: Apico-alveolar realization of /t/ in English (left) and lamino-dental realization of French /t/ (right)

there is sufficient airflow across them. There is energy in the very low frequencies (the so-called voice bar). The passage of airflow is necessary for the vocal folds to vibrate. There is airflow when the transglottal airflow is positive. The transglottal airflow is the difference between intraoral pressure and supraglottal pressure. The intraoral pressures rises during the occlusion phase of the stops and when it is equal to the subglottal pressure, the transglottal is null: no vibration can occur: the vocal folds do not vibrate or stop to vibrate. A positive transglottal pressure is more difficult to maintain when the constriction is far in the back of the tongue because it is difficult to expand the volume between the glottis and the place of constriction, as in the case of the velar consonant. It is much easier to prolong the voicing of the labial consonant /b/ as compared to the velar consonant /g/. Try youself!. Enlargement of the supraglottic cavity volume is necessary to maintien voicing , and it is done via tongue root advancement in the case of /g/, Westbury (1983). As a consequence of the aerodynamic constrain to maintain a positive transglottal airflow, /g/ is most likely to to be missing in the series of voiced stops /b,d,g/. (On the glottal activity and intaoral pressure during stop consonant production, see Warren & Hall (1973). Spanish and French show systematic use of voicing maneuvers compared to English, Solé (2018).

For the stops, the closure in the vocal tract is complete. For the fricatives, the vocal tract is constricted, without complete blockage, and a friction noise is created at the place of the constriction. The pressure behind the constriction rises: therefore the transglottal pressure diminished but to a lesser degree than for the fricatives.

In the case of the voiced fricatives it is however difficult to maintain voicing for articulatory reason. The glottis has to be open in one portion for letting the airflow to go to the supraglottic cavity to create friction noise, on one side and to be loosely close in another portion to insure the presence of vibrations, on the other side. A strong friction is only obtained at the expense of voicing and vice versa. If the strength of the friction noise is favored, the voiced fricative becomes voiceless, and if the voicing is favored, the voiced fricatives become an approximant, Yeou & Maeda (1995). (On voicing in plosives and fricatives, see Maddieson (2013).

What is the main difference in the realisation of the voicing feature for the stops in French and in English? The presence/absence of the voice bar in French is essential for the distinction between voiced/unvoiced stops in French. The long/short VOT (Voice Onset Time, that comprise the duration of the release, the friction and the aspiration noise) is essential for the distinction between voiced/unvoiced stops in English. Presence/absence of voice bar and duration of the VOT are not the only acoustic cues contributing the the distinction between voiced and unvoiced consonants. In determining whether a listener reports hearing 'rapid' or 'rabit' as many as 16 pattern acoustic properties can be counted that may play a role in determining whether a listener reports hearing one of these words rather than the other ,Lisker (1986).

The boundaries between contrasting categories along the continuum of voice onset time vary also from language to language, Lisker & Abramson (1964).

What about sonorization (voicing) of the voiceless consonants and the desonorization (devoicing) of the voiced consonants? Voicing and devoicing is a result of sound assimilation with an adjacent sound of opposite voicing. An unvoiced consonant generally become voiced if followed by an voiced consonant (sonorization voicing). An Voiced consonant generally become unvoiced if followed by an unvoiced consonant (desonorization, devoicing): /bs/ is heard as [ps] in the word "absent" in French.

3) The third criteria is *degree of constriction.*

Oral stops (/p, b, t, d, k, g/) imply complete closure of the vocal tract. The intraoral pressure goes up during occlusion.

The fricatives (/ f, s, ʃ, v, z, ʒ / are produced by a strong narrowing in a more or less constricted zone of the vocal tract, which becomes the production place of a supraglottal friction noise. The friction noise is then filtered mainly by the cavity in front of the constriction point. The friction noises corresponding to /s/ and /z/ are more acute than those corresponding to /ʃ/ and /ʒ/ because the tongue fronting and lip spreading shorten the front cavity, while withdrawing the tongue and/or rounding the lips lengthen it in the case of /ʃ/ and /ʒ/.

4) The fourth criteria is *nasality*.

Figure 7.4 illustrates the configuration of the tongue, the lip and the velum for the production of the nasal stops /m/ and /n/.

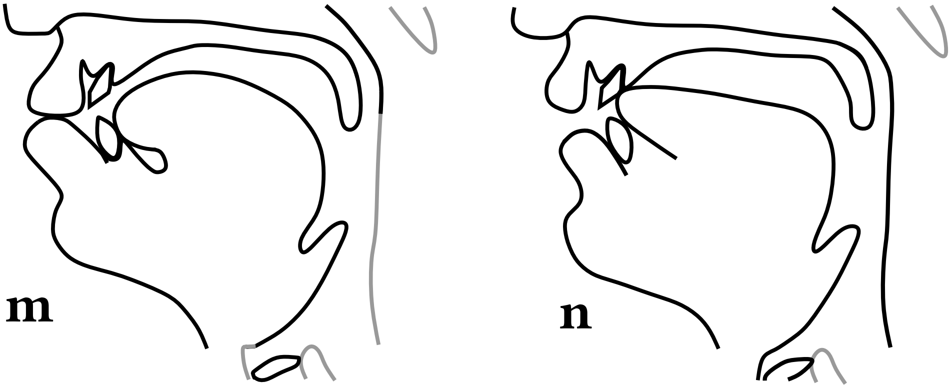


Figure 7.4: Lowered velum in nasal consonants

The nasals /m, n, ɲ/, are produced in the same way as the corresponding voiced stops, except that the pharyngo-nasal passage is open. T he nasal cavity resonates and thus participates in the filtering of the voice source in the oral cavity. Nasals are characterized by the presence of formants and anti-formants. The acoustic characteristics of the nasal consonants are much less complex than those of the nasal vowels, for the analysis of nasal consonants, Fujimura (1962).

The nasals are generally voiced, but voiceless exist, Bhaskararao & Ladefoged (1991).

Not all sounds have pulmonic airflow as their origin. It is possible to create airflow without the involvement of the lungs as in *implosives* (the implosive airflow is due to the closing and lowering of the larynx), *ejectives* (the evasive airflow is due to an elevation of the larynx) and clicks, where the air is trapped between two constrictions in the vocal tract, Ladefoged (2005). In languages where non pulmonic sounds do not have phonemic status, they are used for expressive purposes, such as the dental click in French, to express annoyance.

The consonants are realized differently depending on their position relative to the word stress and to the word boundaries. A consonant may be deleted in certain position: American English speakers generally do not pronounce the [d] in ”handball”. For example, in strong position (i.e. in word-initial position and in pre-stressed position), unvoiced stops are aspirated in English. They are unaspirated if they occurs after /s/ (star). In American English, /t,d/ is generally realized as a glottal stop in post-stressed position or as a flap : butter /ˈbʌtɚ/ [ˈb̥ʌɾɚ] or a nasal-released stop (garden) or as nasal : want to —-> wanna (on the study of medial /t,d/ , Zue & Laferriere (1979).

Chapter 9 (on Prosody) deals with the effect of lexical stress and word boundaries on the realization of the consonants.

8 Some aspects of speech perception

The invention of formant synthesis using the *pattern playback* at Haskins Laboratories at the end of the Second World War 1945 marked the start of scientific studies on the identification of individual sounds. The pattern playback is an electronic device for converting hand-drawn spectrographic representations on a transparent medium into sounds. The research concerned in particular the role of the formant frequencies on the identification of the stop consonants /b/, /d/ and /g/, and their discrimination. ? A typical question for an identification task is the following: (are you hearing /b/, /d/ or /g/?) and for an discrimination task: (are these two sounds corresponding to the same consonant or not?)

Psychoacoustics is a branch of science dealing with the the relationship of human's subjective response to sound waves changing in frequency, duration and intensity. *Psychoacoustics* is mainly concerned with the perception of pure sounds (called *pure* due to the fact that they are generally composed of a single frequency or a few frequencies, unlike speech sounds which are very complex sounds.

At that time of the invention of the pattern playback, thanks to the work done by *psychoacousticians* a lot of knowledge had accumulated on the perception of sounds and on the physiology of the auditory system

- the *audibility zone* of an individual designs the zone between the audibility threshold and the pain threshold; to be perceived by a human ear, sounds must have a frequency greater than 16 Hz and less than 16kHz-20kHz (10kHz in some elderly subjects) and they must have a sufficient intensity (which depends on the frequency). The sounds produced by certain animals, such as ultrasounds and infrasounds, and used to communicate with their peers are not all audible to the human ear;

- the *sensitivity of the ear* varies according to the intensity of each frequency and the maximum sensitivity of the human ear is reached for frequencies between 2000 and 5000 Hz;

-the *perception of frequency and intensity is non-linear*: up to 10kHz, the perceived pitch is linear in relation to frequency, but for higher frequencies, larger changes in frequency are necessary in order to perceive changes in pitch.

- the *temporal and frequency masking effects*: a low tone tends to mask an higher tone occuring simultaneously; one sound can obscure another sound, occurring before or after it;

- the notion of *critical bandwidth*, introduced by Harvey Fletcher in 1933: the ear cannot distinguish two pure tones within the same critical band that occur simultaneously etc.

For a review, (Handel 1993).

The *pattern playback* is at the origin of great discoveries Firstly, despite the fact that all speech sounds and the sounds of nature (often called *noises*) use the same auditory pathway to the brain, it was shown that speech sounds are not treated in the same way than the other sounds at a more or less peripherical levels: the general auditory principles well established by the psychoacousticians could not readily explained a number of phenomena of speech perception.

Secondly, it was also demonstrated that each perceived phoneme corresponds to a very complex acoustic reality, and that its identification corresponds to a very complex mechanism. The experiments with the *pattern playback* illustrates the non-uniqueness of the acoustic cues relative to the place of articulation of the same consonant followed by different vowels. They allowed the discovery of the phenomenon of *categorical perception* , which was believed for some time to be specific to human speech, and led to the formulation of the *motor theory* of speech perception (Alvin Liberman, (Liberman & Mattingly 1985). When the listener is asked to identify the place of articulation of the consonant, while the formant transition varies in a continuous way, he perceived easily and categorically either /b/, /d/ or /g/ (*categorical perception*). The *motor theory* states that the listener is *perceiving* the gesture done by the speaker to produce the sound from the acoustic signal. This theory will be reviewed later in this chapter. After the arguments advanced by proponents and opponents of the *motor theory* , a long quest followed to explain how a listener can form a single invariant percept (such as /t/, for example) from the multiple and variable cues contrasting /b/, /d/ and /g/ contained in the speech signal. There is, as yet, no unanimously accepted theory to explain the invariance of percept in opposition to the variance of the acoustic signal corresponding to the realization of a single phoneme. And Alwin Liberman to conclude that speech corresponds to a special code. Liberman (1996).

The spectrographic representation of the voiceless consonants /p/, /t/ and /k/ allowed to discover other acoustic cues for identifying the *the place of articulation of stops* other than *formant transitions*: in particular the *noise burst* which is not invariant but vary in function of the following vowel, in particular in the case of /k/.

Since the mid-1990s reflections on the understanding of *meaningful utterances* and on the contribution of knowledge of discourse situations have somewhat diverted researchers' attention from the identification of individual sounds and purely psychophysical aspects. As stated flatly by Roman Jakobson, the speaker speaks to be heard and understood by the listener, (Jakobson et al. 1978): and the listener tries above all to make sense of what he hears. His goal is to interpret a message rather than decode a sequence of phonemes. In fact, to understand the meaning of an utterance, it is only necessary for him to identify *most of the words* that compose the message.

The research questions are the following:

- How and when does the listener *segment* and *identify* successive words in spontaneous speech in the continuous flow of speech ?

- How does the listener use the *multitude of cues of all kinds* (syntactics, pragmatics, semantics, visual information, etc.) to decode speech, sometimes in the presence of noise? In other words, What is the part, in the instantaneous understanding of an oral message, of the acoustic information provided by the signal itself or the speaker's face (inductive information: *bottom-up*), and the part of the discourse context (deductive information: *top-down*)?

- How are all of the words stored in the *mental lexicon* represented: is each word represented in the brain as an ordered set of *phonemes* , or as a set of *distinctive features*? or as one abstract *prototype*? or as a collection of *detailed episodic traces* from a previous listening of the same word (i.e. multiple *copies* accumulated in our long-term memory, accounting for all of his experiences)?

- What is the access strategy of the mental lexicon where the words are stored? are they accessed by their beginning, by their stressed syllables?

Despite the efforts made in this area, our understanding of the phenomena related to speech perception and message comprehension still falls far short of the knowledge gained on production. Several strategies of understanding speech have been proposed and there is no definitive answer.

Firstly, several strategies could coexist and one of them could dominate, depending on the circumstances (for example in a very noisy environment). Each experiment highlights the ability of listeners to use a particular strategy or clue in a particular task, and a different particular strategy or clue in another task, but they do not provide much information on the actual strategy (or combination of strategies) used by listeners on a daily basis in real time.

Secondly, controlled experimentation on the decoding of truly spontaneous speech is difficult because the delayed judgment of the listener in laboratory conditions is influenced by a large number of parameters such as his *motivation*, his *growing familiarity* with the requested task (hence the answers to the same question may change over the test), *the clarity of stimuli, and of the voice of the speaker(s)* and/or with the *topic*.

Thirdly, acoustic characteristics of messages and expectations about their content from previous knowledge of the context have a known influence on the understanding of everyday messages, which is difficult to estimate.

As a consequence, experiments are often hardly reduplicable.

All sounds arrive in the auditory areas of the cortex in the form of nerve impulses and it is difficult to estimate the exact difference between the early treatments of human sounds and the noises of nature.

Figure 8.1 illustrates the ear diagram and the auditory field.

The *mechanical vibrations* of air particles that make the sound wave are sensed at the level of the auricle of the ear and transferred along the *external auditory canal* to the elastic membrane of the *eardrum* which is consequently set into vibration. The auditory canal amplifies frequencies that are close to 3500 Hz as they pass through it. The vibrations are then transmitted to the *middle ear* where a chain of three auditory *ossicles* — the hammer, the anvil and the stirrup — increase their amplitude via a lever action and cause the basilar membrane to vibrate. The vibrations are then converted into *electrochemical nerve impulses* by the approximately 25,000 hair cells distributed at the level of the cochlea of the inner ear. Each hair cell *vibrates* within a certain frequency zone, which depends on its position on the cochlea. These impulses reach the brain through the *auditory nerve.*

This zone of *perceptibility* of sounds by humans includes the area of the lowest frequencies produced by the vocal folds (around 75 Hz) and the highest frequencies as in the most acute sounds, such as the consonant /s/, produced by a human vocal tract (up to 6000 Hz is sufficient to detect the differences in the distribution of energy in the high frequencies between /ti and /k/ friction noise, diffuse and rising for /ti/ and compact for /ki/. Thus, the ear carries out a form of *frequency analysis* of the signal, as a filter bank, in the manner of a spectrogram (Fourier analysis) but the analysis is not linear: the low frequencies are analyzed with more finesse than the high frequencies. Conversely, the temporal resolution is better for high frequencies. In addition, there are phenomena of *frequential and temporal masking* ,that is, at a given instant, certain frequency components mask others and are not audible. In general, low frequencies tend to mask high frequencies (so-called frequential masking). However, the high frequencies can mask the perception of low frequencies, as the energy concentration due to the grouping

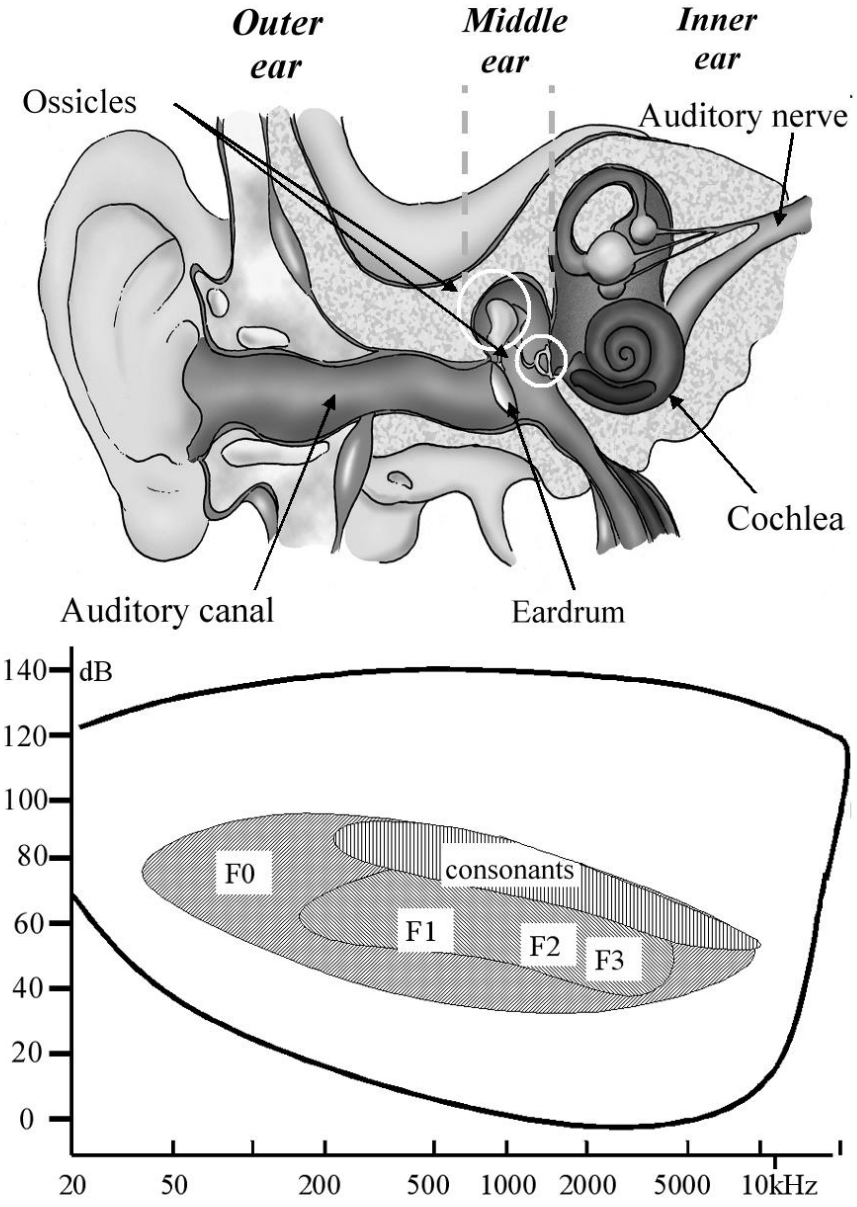


Figure 8.1: . — Ear diagram (above); the auditory field (below) with a rougth indication of the regions where H_0, F_1, HFtextsubscript2 and H_3, and consequently the vowel formants, and the the regions covered by the consonants

(F_3-F_4) around 3000 Hz masks the perception of the F_2 of the vowel /i/, as mentioned before. Similarly, a low-intensity sound may be masked by a louder sound that precedes or follows it — this is the so-called temporal masking. For example, the nasal consonant in coda must be of sufficient duration in order not to be perceptually masked by the preceding vowel, which is always more intense

Some remarks on speech perception : the perception of loudness and stress, the perception of vowels, and the perception of consonants. The perception of intonation is left apart, but some elements are discussed in Chapter 6.

1. *On the perception of loudness and stress:*

   Loudness plays a large role in speech, for the perception of prominence (stress), Bolinger (1982) and rhythm. The intensity of the *sounds/noises of nature* (not produced by an human vocal tract) tends to be estimated by the listener according to their real *acoustic intensity*. A sound with higher intensity is perceived as louder. Psychoalinguistic experiments thus show very clearly that one does not perceive the intensity of the *sounds of speech* in the same way as other sounds or the sounds/noises of nature: for example, the subjective intensity of a speech sound (loudness) is judged by the listener in relation to the *estimated vocal effort* made by the speaker, Lehiste & Peterson (1959). The overall intensity of a vowel is determined mainly by the *first vowel formant frequency*. The open vowel /a/ (which an high F_1) has an *intrinsic acoustic intensity* much higher than that of the vowel /i/ (which an low F1), but the vowel /a/ will be perceived as having an intensity equal to that of the vowel /i/ if the listener thinks that the speaker has made the same articulatory effort to pronounce the two vowels, Lehiste (1970), page 235.

   - A syllable has to be perceived as relatively more loud than the surrounding syllables to be perceived as stressed (prominent). Loudness increases with *duration* (temporal integration of loudness, i.e. the level difference between equally loud short and long tone) and varies with *frequencies.* and amplitude for the perception of prominence within a syllable is very complex.

   - The fundamental frequency and amplitude contrast between stressed and unstressed syllables is significantly less during soft speech as compared to normal speech, suggesting that duration is the predominant feature of stress during soft speech Mcclean & Tiffany (1973).

   - Furthermore deafness the perception of which syllable is the most prominent in a word depends on the listener's native language: French subjects exhibit much more difficulties in discriminating non-words that differ only in the location of stress, contrarily to Spanish listeners (the position of stress in a word is distinctive in Spanish, but not in French).

As a consequence, the interaction between too many parameters (intensity, duration, Fo, rate of speech, speaking level, the native language of the listener) it rather impossible to estimate quantitatively the degree of loudness that will be perceived by the listener.

2. *On the identification of vowels:*

   The cues used by the ear to identify a phoneme (a vowel or a consonant) take into account *the expected target* for the phoneme and *the possible deviation* due to the phonetic context in which it occurs, i.e. the surrounding phonemes, its position relative to the word stress and boundaries and the speaker. For example, the same syllable may be perceived differently depending on whether the carrier phrase is produced by a man or a woman. In the case of a man's voice, the listener expects relatively low frequencies and tends to overestimate the formant values. He adapts his expectations to the frequency characteristics of the perceived voice, Ladefoged & Broadbent (1957).

   - The timber of the *oral monophtongs vowels* are entirely descripted by the frequencies of their formants, taken in the middle of the vowels, as proved by formant synthesis. The relative formant amplitudes of the formant, in modal phonation, can be largely computed from formant frequencies and are somehow redundant information with frequencies are vice versa. For example, for the neutral vowel, with F_1 at 500 Hz, F_2 at 1500 Hz, F_3 at 2500 Hz, F_4 at 3500 Hz. F_1 amplitude is expected to be the largest amplitude, higher thant F_2 amplitude and F_4 amplitude the lowest. For a male speaker, F_1 and F_2 are expected to be under 2200 Hz and F_1 between 250 Hz and below 1200 Hz. Due to acoustic laws, the proximity of formants causes an increase in their mutual amplitude, due to the laws of acoustics, and this increase creates a perceptual salience of the frequency zone corresponding to the regrouped formants , Fant (1960a).

   - The frequencies of the *first two formants* (Chapter 6) are sufficient for synthesizing the *grave oral vowels*, of which the F_2 is low (below 1200 Hz) and thus not very far from F_1. The ear integrates the two formants and perceives only one peak (the 'center of gravity'effect, Chistovich & Lublinskaya (1979). Even only *one formant* is sufficient to accurately yield the timber of the dark, focal, labio-back oral vowels (such as the realization of the cardinal /u/), where F_1 and F_2 are close together and thus of high amplitude, masking higher formants (such as F_3 and F_4).

- The formants higher than F_2 (i.e. F_3 and F_4) influence the perception of *acute vowels* because they are of greater amplitude than the higher formants (F_3 and F_4) of grave vowels and are known to have a perceptual weight. If a vowel of the /i/ type synthesized with the first four formants, respectively 255, 2065, 2960, and 3400 Hz, is presented to listeners, and if the value of F_1 is set at 255 Hz and if they are asked to adjust the value of only one formant in order to obtain a timber that is as close as possible to the synthetic vowel with four formants, the listeners will adjust this value to approximately 3210 Hz, that is to say a value situated between F_3 and F_4 which are closed together. This resulting formant (the *effective formant*) is called *F two prime* or F_2'. A representation of the required value like F_1 / F_2' is better than F_1 / F_2 or F_1 / (F_2 / F_1) for the acute vowels, Fant & Risberg (1963). As shown by formant synthesis, the timber of the acute vowels obtained with F_1 and F_2' alone, although close to the original vowels, is not exactly the same as the timber made by all the formants involved: the way that formants higher than F_3 are distributed plays a role.

- What about the nasal vowels? *Lowering the amplitude of the first formant* relatively to the second formants leads to a higher degree of perceptible vowel nasalization Delattre et al. (1952). F_1 and F_2 are not sufficient to account for the timbers of the nasal vowels and the vowels with voice quality contrasts. Listening is mostly guided by expectations. Nasal vowels are easy to be recognized on spectrograms because there is no the expected correlation between the spectral peak amplitudes and their frequencies and the position and number of spectral peaks as compared to oral vowels. See the collection of articles on nasality, Anderson et al. (2014).

- If F_2' is taken as a reference point, the vowels in French can be listed in terms of their F_2' in descending order as follows: /i/ (F_2' around 3000 Hz, close to F_3), /e/ and /ɛ/ (F_2' slightly above 1800 Hz), /ø/ and /œ/ (F_2' around 1500 Hz), and /ɑ/, /ɔ/, /o/, /u/ (F_2' below 1200 Hz and close to F_2).

/i/, /y/, /u/ and /ɑ/ are focal vocals. Focal vowels are characterized by a dense concentration of energy in a reduced zone of frequencies created by the proximity between two formants, F^n and F^{n+1}. To be perceived as French-like, the *French vowel /u/* must be of the focal type (F_1-F_2) with F_2 as low as possible. This vowel is the darkest vowel that a human vocal tract can produce, with its two main resonances below 1000 Hz, and perceived as such. The *French vowel /a/* is also of the focal type (F_1-F_2), like /u/, but with the highest possible values of F_1 around 1000 Hz. The *French vowel /i/* is of the (F_3-F_4) type, sometimes of the (F_4-F_5) type, is the most acute vowel (F_2' around 3000-3200 Hz) that a human vocal tract can produce, with F_3 as high as possible, as confirmed by experiments in articulatory synthesis. The whole of the vocal tract deforms in order to achieve the intended acoustic goal and compensations are possible, between the articulators. For example, raising the larynx during the production of /i/ allows to keep F2 high. These compensations indicate the primacy of a perceptual rather than an articulatory target for sounds, as already mentioned.

Vowels may be difficult to discriminate and to identify. Unlike the consonants which are identifiable by a limited number of items (for example, we perceive /p/, /t/ or /k/) the discrimination between vowels when F1, F2 and F3 are continuously varied is more problematic. The perceptual confusions between vowels are generally based on their acoustic resemblance. The pairs of vowels in which the two members differ in degree of aperture /u-o/, /o-ɔ/, /ɔ-ɑ/, /e-ɛ/, /ø-œ/ are expected to differ mainly in F_1 value (lower F_1 frequency is expected for the first member of the pair). However, the two members of one pair may be realized for certain speakers with the same F_1 frequency but lower F_2 frequency for the first member of the pair. In general, a speaker is consistent in his strategy to maintain the contrast and there is no perceptual confusion between acoustically close phonemes. Perceptual confusion may happen due to coarticulation. The consonant context can provoke an articulatory anteriorization or posteriorization of the vowels, which may lead to the perceptual confusion on the antero-posterior axis in the listener: do we pronounce *reblochon* [ʁœbloʃɔ̃] ɔn] or *roblochon* [ʁobloʃɔ̃] ɔn] “a type of cheese”? Most French speakers do not know. The influence of the phonetic context on the articulation of vowels and consonants has caused many phonetic changes, which have gradually separated the spelling from pronunciation, even if the latter, during its elaboration, has partially reflected the phonemic system. The French word *fait* /fɛ/ was formerly pronounced [fa^{i}t]. The spelling in French mainly reflects how the words were pronounced in the 13th century.

3. On the identification of consonants

   - The acoustic cues for the *place of articulation of stops* are, as previously mentioned, essentially the frequency characteristics of the *noise burst* and the *formant transitions*.

For what concerns the *burst*, identification experiments of the place of articulation of stop consonants (/p/, /t/ or /k/) in synthetic stimuli, researchers at Haskins Laboratories have already shown that a *noise burst* at the same frequency zone played in front of a set of different vowels synthesized with only two formants could give the perceptual impression of different consonants, depending on the formant frequencies of the vowel that follows. When the noise burst has a high frequency, [t] is readily perceived, regardless of the following vowel. If it is of low frequency, [p] is perceived. The perception of [k] depends in most cases on the frequency position of the noise burst with *respect to the* F_2 (better said F_2') of the following vowel: if the noise burst is at the F_2 level of the vowel as in low-F_2 vowels, or at a frequency level between F_2 and F_3 for the acoustically average-frequency vowels (with F_2 at 1500 Hz), or at a much higher frequency as in vowels with higher F_2, it is [k] that is perceived. The identification of /k/ is therefore contextual and requires a special relationship between the frequency value of the released noise of the stop consonant and that of the F_2 (or even better the F_2', see above) of the following vowel. This explains the articulatory adjustment often observed in languages where /k/ is velar when followed by a rounded back vowel but palatal when followed by a front vowel. This articulation adjustment corresponds to a perceptual necessity. In many other cases, the articulation adjustment of a phoneme in function of its phonetic environment seems to correspond to a need for articulatory simplification, by lessening the articulatory distance between two successive phonemes.

- For what concerns the *formant transitions*, previous studies with synthetic speech (pattern playback) of researchers from Haskins have also shown, in regards to synthetic speech with only *two formants* , that F_2 transition variations were sufficient in synthesis to distinguish between [p], [t] and [k], without it being necessary to reproduce a noise corresponding to the noise burst. The locus, for the second formant, corresponds to the relatively fixed place of production of the consonant/ For the voiced stops, /b/, /d/, and /g/, there appears to be a locus for d at 1800 Hz and for b at 720 Hz, but no /g/ locus was found, Delattre et al. (1955).

Note that in natural speech, with *the noise burst and formant transitions* are necessarily present: the respective weight of the noise burst and transitions depend on the intrinsic nature of the consonants and the following vowel. For example, the F_2 transition is ineffective in distinguishing between /ti/ and /ki/: only the spectral distribution of the noise burst in the high frequencies counts, which is more compact in the case of /k/ than in /t/. F_2 transition are the main for distinguishing between /pa/, /ta/ and /ka/ and the noise is not necessary: F_2 transition. In general, the frequency height of the noise, its intensity and compactness, as well as the formant transitions, contribute, in varying degrees, to the identification of the place of articulation. Some combinations are more difficult to identify than others. /ti/ and /ki/ are more difficult to distinguish than /pa/ and /ka/ by children, the hearing impaired, the elderly (presbyacusis), on the telephone and by the spectrogram reader.

-The main acoustic cue (friction noise or formant transition?) for the identification of the place of articulation of the *fricatives* depends on the intensity of the friction noise.

For /s/, /z/, /ʃ/ and /ʒ/ which have high intensity noise (*sibilant fricatives*), the main acoustic cue is the position of the concentration of energy: friction noise above 4000 Hz or > F_4) for /s/, /z/, and [ʃ] (between 2000 and 4000 Hz) for /ʃ/ and /ʒ/. The transitions into the next vowel play a minor role.

In the case for bilabial /ɸ-β/ or labiodental /f-v/, which have weak intensity noise (i.e. *non-sibilant fricatives*), the main acoustic cue is the position of the concentration of energy, see Harris (1954) on the cues for the identification of the fricatives of American English.

Languages prefer high-intensity fricatives in their inventory such as /s/ and /ʃ/ for their intrinsic characteristics, and avoid pairs of phonemes difficult to differentiate, such as /ɸ-f/ or /β–v/ with a bilabial-labiodental contrast.

What about the *approximants*? A dynamic cue, such as the speed of transitions, is essential for the distinction between [b] and [w]: the transitions are quasi-identical but slower in the case of [w] and there is no released burst. [w, j, l, r] are recognized by their own formants and the transitions from their own formants toward the surrounding vowels. The presence of traces of nasalization at the release of the stop and at the beginning of the next vowel is one of the cues for distinguishing between [m] and [b], [n] and [d] (Stevens). The distinction between [m] and [n] in continuous speech is ensured mainly by the transitions of F_2, not by the acoustic characteristics of the spectra durant the nasals.

- What about *voicing*? As mentioned before in Chapter 7, there is textitlarge catalogue of acoustic features signaling /b/ versus /p/ in trochees, see Lisker (1986).

The weight of each cue for deciding whether or not a stop is voiced or voiceless vary *from language to language*. Different languages prioritize different cues to distinguish between voiceless and voiced stops (for a acoustic cross-language study of voicing in initial stops acoustic study, Leigh Lisker, Arthur Abramson, Lisker & Abramson (1964).

The weight of each cue also depends also on the context, mainly the position of the consonant within the syllable (onset or coda), and relatively to the word stress and the word boundaries .

Let us take an example concerning the word initial stop consonants. For native of English, the main voicing cue is the delay time of voicing, called VOT (Voice Onset Time), of which the interpretation is language specific , for a cross-language study of voicing in initial stops, Lisker & Abramson (1964). The same sound will be perceived as /b/ by English listeners if voicing (related to the vocal fold vibrations) begins very quickly after the release (less than 30 ms), and as /p/ if the delay is greater than 40 ms. For a French listener, the sound is perceived as /b/ if the vibrations begin before the release: an English /b/, when devoiced – such as at the beginning of a word – may be perceived as /p/ by a French speaker.

Consonants and vowels display strong acoustic similarities in terms of their *F-pattern* if they are pronounced with a similar place of constriction in the vocal tract. The constriction is less narrow in vowels than approximants, and less narrow in approximants than obstruent consonants. As a consequence, the formant values are the most extreme in obstruents, but the F-pattern remains similar for all phonemes sharing a similar place of articulation, due to the laws of acoustics. The timbers of the approximants [w, j, ɥ] are perceptually similar to those of the vowels [u, i, y] and their F-patterns are very much alike too, but their F_1 is lower due to a more extreme closing and thus their intensity is reduced. There is also timber resemblance between the French [ʁ] in the context of the the back vowel [ɑ] (1000 Hz), between the dark [ɫ] (as in the English word *film*) and the back vowels [u] or [o]. Accordingly, the vowels /i/, /ɑ/, /u/, /y/, /u-o/ correspond to /j/, /ʁ/, /w/, /ɥ/, /ɫ/ in terms of F-pattern. For example, if the portion correspond to [ʁ] is extracted from the syllable /ʁɑʁ/, the vowel /ɑ/ is heard. This perceptual proximity between the dark [ɫ] and [u-o] is at the origin of its transformation into [u] in syllable-coda position in the passage from

Latin to French: *soldus* [soldus] > *soldu* [sołdu] > *sold* > *soud* > sou [su]. When the French /i/ (F_2' higher than 3000 Hz) gets devoiced (due to too much narrowing of the vocal tract), it gives rise to an identified sound resembling the German palatal fricative [ç] (pronunciation [ɥiç] noted *uiche* from *oui /wi/* in everyday French). /i/ and /ç/ share a similar F-pattern and the friction of /ç/ is closed to devoiced /i/. (See Fant (1960a) and Vaissière (2007) on the F-pattern). Note that the phonological features currently used in literature for characterizing the phonemes unfortunately fail to stress the similarities between a vowel, a semi-vowel and a consonant sharing a similar place of constriction.

4. Some models and theories on the identification of the phonemes.

   There are several theories dealing with the identification of phonemes. Some theories refer to the articulation of sounds, other to the intrinsic acoustic properties of the phonemes.

   - *Categorical perception* refers to the articulation of sounds. It has until now occupied an important place in reflections on the phoneme because it has long been regarded as specific to the perception of speech by humans. We now know that we can perceive noises and colors in a categorical way, and that animals can likewise employ categorical perception. In this regard, categorical perception is not specific to the perception of speech by humans. Categorical perception implies that identification precedes differentiation, that is, the comparative judgment of the timbers. *Categorical perception* has been found mainly for the identification of the consonants, because the choice offered to the listener is limited to the consonant labels existing in his own native language: a consonant is generally identified as /p/, /t/ or /k/ and not something in between. For what concerns the vowel, it is less clear, the perception is more continuous. One may hesitate between perceiving an /u/ and an /o/: the vowel may have an F1 to high to be recognized as an /u/ but an F2 low enough to be recognized as an /u/.

   -*The motor theory of speech perception* [1] advances that in order to identify a sound, the listener interprets what he perceives in terms of articulatory gestures. The categorical perception of the place of articulation of consonants has been interpreted as favorable to the motor theory. The phonetic realization of the consonants /p/, /t/ and /k/ requires, in fact, well-differentiated articulators: for example, the lips for /p/, the tip or edges of

[1] Liberman *et al.*, on the perception of the speech code, 1967.

the tongue blade for /t/, and the tongue body for /k/. In identifying the place of articulation of these consonants the speaker *sees the gesture*: he would refer to the way in which he himself would have produced these sounds, and which articulator he will use, so that there would be clear perceptual boundaries based on precise articulatory criteria (lips = /p, b, m/, tongue blade = /t, d, n/, tongue body = /k,g/. Nevertheless, newborns perceive certain contrasts of consonants in a categorical way, although they have never pronounced these sounds, which casts doubt on the role played by the reference to production — unless we imagine that correspondence is included in their genes! On the other hand, for vowels the tongue can adopt an indefinite number of positions along both the vertical and the anterior-posterior axes, and there may be compensatory gestures and their perception would not be categorical for this reason. On On categorical perception and the motor theory, Liberman et al. (1967).

- Other theories do not refer to the articulation of sounds but to their intrinsic acoustic properties.

- We have already mentioned Stevens's *theory of invariance* in Chapter 5, Stevens & Blumstein (1981). According to this theory, the phonemes are directly recognized from certain invariant acoustic properties without reference to articulation. Some sounds, such as the vowel /i/ and the consonants /s/ and /ʃ/, have intrinsic acoustic realizations that are relatively invariant, others less, such as the vowels /o/, /ɔ/, /e/ and /ɛ/. Training in spectrogram reading allows the decoding, without great difficulty, of carefully articulated meaningless words, both in French and in other languages. This ability suggests the presence of at least relative acoustic invariance. The extent of acoustic variability of the phonetic realization of the phonemes seems to have been somewhat overestimated, especially by psycholinguists.

-Fant prefers the term *relative invariance.* The noise burst level of velars, with a noise more compact than that of labials and dentals, is not strictly invariant, but has to be interpreted according to the F_2 value of the vowel or even better according to the effective F_2' value .

- Some recent (*episodic* or *exemplar*) models hypothesize that the acoustic image of each word heard by the listener is stored as such in his mental lexicon, and that memory is virtually unlimited. This conception amounts to putting the abstract notions of features and phonemes on which phonology is constructed in the background. It is currently admitted that there are differences in *quality* between different realizations of the same phoneme.

Some realizations seem to be good representatives of the category of that phoneme in a given language, others less so. The listener is able to judge whether a stimulus is a more or less remote exemplar of what he considers the ideal prototype for a particular vowel. The increase of the reaction time set by the listener to locate a given phoneme in continuous speech generally indicates that the stimulus to be identified is not prototypical, even if correctly identified. For this reason, new models of speech perception are oriented towards an exemplar and probabilistic framework.

In communication, each phoneme of each word is not necessarily identified before understanding the word or the overall message. The perception and comprehension of continuous speech involve *central mechanisms*: words and whole utterances are recognized by an interaction between acoustic cues decoded from the signal on the one hand and the mental lexicon (bottom-up information) as well as syntactic, semantic, and contextual knowledge on the other (top-down information). The speaker *adapts his way of speaking* to the context, in order to be understood. In this regard, he will allow himself a certain articulatory laziness and will even omit certain phonemes (or words) if he is convinced that he is nevertheless understood by his interlocutor. Under certain circumstances, he will allow himself to pronounce [ʃpa], instead of [ʒənələsɛpa], *je ne le sais pas*, "I don't know it", as mentioned before. On the other hand, he will make a particularly strong articulatory effort when he talks to a child or a foreigner or in noisy conditions. Some speakers constantly speak in a relaxed manner, leaving their listeners to make sense of what they say by using the context. Others, like teachers, speech professionals or mothers of young children tend to hyper-articulate, thus producing better representatives or exemplars of each phoneme and word. Moreover, the sound material that precedes the sound to be recognized influences the judgment of the listener. The acoustic signal corresponding to the This adaptation can be done very quickly, and a listener can *adapt* rapidly to the recurring defects of pronunciation of his interlocutor.

What to conclude?

- *Speech perception has not yet revealed all of its secrets* .

  The listener may perceive sounds (or silences) that, in fact, are not present in the signal.

  For example, in continuous speech, the listener may perceive a *pause* between two words in the absence of any silence: a rise of the fundamental frequency (the so-called continuation rise) or a lengthening of the word or phrase-final rhyme (the so-called lengthening) may, in French, give the impression of a pause, Karcevski (1931).

There are also phenomena of phonemic restoration. See Warren & Sherman (1974), on phonemic restorations based on subsequent context. That is, when a sound segment corresponding to the realization of a given phoneme in an utterance is replaced by a noise, if the utterance makes sense, the listener understands it without any effort despite experiencing difficulty in specifying which speech sound segment is missing in the utterance. In fact, he *hears* all of the sounds even if they are not all present in the spoken chain.

- Certain aspects of perception that have been thought to be specific to human perception, such as categorical perception, have been found to be due to the general properties of the auditory system of primates. It would seem, however, that the *constitution of the prototypes of sounds* is specific to humans. The intensive exposure of an ape to language sounds does not seem to lead to a psychoacoustic reorganization around phoneme prototypes particular to language, as is the case in the human baby.

- Recent research on animals, using cerebral imaging techniques, shows that they react differently to the sounds produced by their peers and other species, suggesting the existence of *biologically specialized mechanisms* to treat the sounds produced by the same species. Animals would therefore lack the mechanisms necessary to treat human speech, but they have detectors adapted to the survival of their species. These mechanisms may be at a relatively peripheral level in the hearing chain.

- There is a certain correlation, maybe as a matter of simple coincidence, between the operating modes of computer tools and the types of models successively developed by phoneticians, phonologists and psycholinguists.

  Models of binary features (supported by the information theory of Shannon), well-adapted to the sequential processing of information by computers of the time (mid-20th century, when the memory capacities of the computer were very limited), gave way to *parallel processing models* precisely at the time when the computer was able to perform such parallel processing.

  The *exemplar models*, currently very popular, are based on the idea that the brain has a very extensive stock of heard occurrences. This is the operation

> mode that evokes that of object-oriented programming and the very large memory capacities of current computers.
>
> *deep learning* are assumed to present the solution for every problem.
>
> This parallelism between technological advances in the computer field and the succession of theories on speech perception is at least surprising, perhaps even worrying.

For introductory readings on auditory phonetics ad perception of speech, we recommend

Antoine Cohen & Sibout G. Nooteboom. 1975. *Structure and process in speech perception.* Springer, (Handel 1993). David B. Pisoni & Robert E. Remez. 2008. *The handbook of speech perception.* John Wiley & Sons. (Pisoni & Remez 2008).

For the perception of Fo and intonation: (Collier & 't Hart 1975), (Studdert-Kennedy & Hadding 1973), (Pierrehumbert 1979), (Hart et al. 2006), (Vaissière 2008).

Journals on auditory phonetics, perception: Ear and Hearing, Journal of Speech, Language and Hearing Research, Perception and Psychophysics, Speech Hearing Research

9 Prosody

The term *prosody* is difficult to define. It has traditionally referred to the study of the quantitative value of vowels (vocalic length) and rhymes in versification. Today its scope has expanded, and it refers to all aspects of speech not related to segment identification, including lexical stress, intonation, and rhythm in particular.

Early conclusions on *prosody* were based on intuition and perception, and not on measurements. As early as the 1930s the linguists of the Prague Linguistic Circle (Vilém Mathesius, Serge Karcevski) pointed out the demarcative function of prosody. They had brought to light the fact that a division of the flow of speech such as the audible division of the sentence into *theme* (new element) and *rheme* (an element already known), an opposition of which the first formulations date back to antiquity, is governed by *pragmatic* factors. In other terms, the main boundary in an utterance is governed by *pragmatics*, not by *syntax*: there is a preference of placing the main boundary between theme and rheme rather than between subject and verb. The same period saw the emergence of a number of works on the teaching of English prosody. Researchers in several languages (English, French, Spanish) noted very early, in listening, the existence of different degrees of boundaries between the words within the sentence: major boundary and minor boundaries, up to 6 'tonemas' for Navarro), opposing an audible final fall at the end of the sentence and audible non-final rises of different sizes at the major boundaries inside the sentence. For French, the perceptual predominance of the final rise at the end of the *sense groups* (a sense-group contains of one or more semantically related lexical words, such as a adjective and the following noun) was noticed in the early 20th century Grammont (1920), Coustenoble & Armstrong (1934), and by Pierre Delattre in the mid-20th century, Delattre (1966c). It was assumed that the regrouping of two or more words into a sense group into French was governed by *semantics*, but as noticed already by Tomas Navarro for Spanish (Tomás (1974), *the maximum number of syllables* in each sense-group plays also a role: the "groupo melodico' in Spanish tends to be seven to eight syllables long. Maurice Grammont noticed the effect of *speaker free choice* for dividing the sentence into more or less sense-groups, and *the speech rate* on the number of sense-groups Grammont (1920): the number of sense-groups tend to decrease as the speech rate increases. Since the 1960s, with Pierre Delattre, instrumental studies have begun to explore the relationship between perceptual impressions and acoustic measurements in the domain of prosody. Instrumental studies confirm the impressionistic reflection of the ancients and the influence of syntactic and pragmatic factors, the free choice of the speaker, the effect of rate, etc.

Generative grammar and the emerging needs of speech synthesis in the 1970s focused on the links between *syntax* and prosody, leaving apart performance factors such as the pragmatic context, the free choice of the speaker and the effect of rate, etc. Most of the studies were done on isolated sentences, read in a ”neutral manner”.

In this regard, various degrees of *boundaries* were confirmed in French and the essentially demarcative function of prosody was highlighted. At least, 5 degrees of demarcation were found in isolated sentence: word boundary, sense-group boundary, main boundary ending either with a sharp rise followed by a non final pause, and final utterance boundary ending with a fall. The literature on English prosody concentrates not on the notion of *boundary* but on the notion of *stress* and on the relationship between different degrees of *stress* and the syntactic structure. The notion of *boundary* is intuitive for a native French speaker, but note that said native has no intuition of what the *word stress* or *lexical stress* may mean, a notion that is intuitive for a native of English, Spanish or Italian. A native of French tends to associate the notion of *word stress* with the notion of ”accent d'insistance”, ”accent emphatique”, generally located at the word beginning, at least for long words.

The work on speech synthesis from written text in various European languages has highlighted the existence of a large number of acoustic marks of prosodic structuring into constituents that are similar but not equivalent to grammatical constituents. Each of the constituent is marked by a number of specific acoustic marks. The acoustic marks are a) the position and length of pauses, b) the fundamental frequency curve/contour, c) the relative duration of the syllables d) the relative intensity of the syllables, e) the allophonic realisation of the consonants and the degree of reduction of the timber of the vowels depending on their position relative to *stress* and/or *boundary*, f) and sometimes the vowel quality.

The exact number of *prosodic constituents* is still a matter of controversy. The prosodic constituents include at least at the phonetic level a) the prosodic *paragraph* (Ilse Lehiste, Lehiste (1975),), b) the prosodic *utterance* (called also sentence), c) the *intonation group* (called also major group, intonational group, breath group, etc.), d) the *phonological syntagma* (more or less equivalent to minor group, accent group or sense group), e) the *prosodic word*, f) the *foot*, g) and the *syllable (*or the *mora).*

The relationship between different degrees of *stress* in English and of *boundaries* in French, on the one hand, and the syntactic structure of sentences on the other hand, was established up to a certain extent for both languages. For the purposes of synthesis of isolated sentences in French, as early as the 1970s, an IBM-France team showed that it was possible to generate an acceptable prosody (Fo contour and length of the syllables) from the syntactic structure alone and the magic number seven to to divide constituents exceeding 7 syllables (Jacqueline Vaissière, Vaissiére (1971).

The clear relationship between prosody and *syntax*, and the relationship between prosody and *pragmatics* are only one aspect of prosody, which moreover fulfills multiple functions. Dwight Bolinger insisted on the fact the accent is predictable only if you are a mind reader Bolinger (1972),. Ivan Fónagy, a precursor, concentrated his research on the expression of *attitudes* and *emotions* by prosody, apart from any syntactic and pragmatic consideration Fónagy (1983)). Further advances in vocal technologies and new orientations in linguistics have drawn the attention of researchers from the prosody of isolated read sentences to prosodic factors in spontaneous speech, in real situations such as dialogues, where the relations between syntax and prosody are less obvious, and where other functions of prosody can dominate. Recent years have witnessed an explosion of research studies related to prosody: prosody and discourse, prosody and the personality of the speaker, the expression of attitudes and emotions, and dialectal and intercultural differences. The face and the whole body may assist prosody or be assisted by prosody in fulfilling such functions. The simultaneous synthesis of voice and facial gestures (the so-called *talking heads*) is also a very active research area. The movements of the eyebrows accompany the expression of astonishment and surprise.

The term *prosody* is a difficult notion to define.

- From an *acoustic point of view*, as mentioned before, it corresponds to pauses, to the fundamental frequency and subglottal pressure variations, voice quality variations (i.e. variations in the vibration mode of the vocal folds), variations in the duration and physical intensity of sounds, and allophonic variations when they are not directly explainable in terms of the characteristics of the sequence of phonemes in the utterance. The modifications of these acoustic parameters sometimes involve the participation of all of the so-called *speech organs* at the glottal, subglottal, and supraglottal levels (such as in the expression of emotions) or may only concern the subglottal and/or supraglottal levels. A decrease in the speech rate, a change in the vibration mode of the vocal folds, an increase in the intensity of the expulsion of air from the lungs, the speed, and the force and the precision of the gestures of the tongue and the lips can carry information about the speaker's involvement in relation to what he says. The more or less subtle deviation of these parameters from the values expected for the same utterance that would have been pronounced in a neutral way conveys information. For example, the *tone* of the voice will be perceived as sad and melancholic, or even joyful and playful.

- From another point of view, prosody can be defined by its *functions*: lexical, demarcating, pragmatic, behavioral, emotional, identifying, and stylistic functions.

- From a *linguistic* point of view, prosody is often described as the sum of the phenomena of lexical stress and intonation, including often performance factors, as well as rhythm. Intonation, like lexical and grammatical stress, is an abstract category, although it is often and abusively identified with one of the parameters by which it is realized, in particular the fundamental frequency (and the perceptual notion of pitch). Intonation designates a) a discrete linguistic system of utterance structuring (a demarcating function, determined by both syntax,pragmatics and length of the constituents), b) a discrete system for expressing modalities (answer, question, order) and c) a non-discrete system for expressing nuances of meaning, social attitudes and primary emotions (behavioral and emotive functions).

All of these definitions are acceptable, none of them are satisfying, and care must be taken not to mix the different functions in a description of the facts (for a review, Fónagy (1989).

Prosody seems to be the child's first language. The newborn is sensitive to the rhythm of his mother tongue. The baby, like domestic animals, is very sensitive to the emotional clues conveyed by the voices of the people around him, as well as to the prosody of his mother tongue. He imitates very early the way his family speaks. The French baby babbles with more frequently rising melodies than the Japanese baby and with a sharper final lengthening (for a review, de Boysson-Bardies (1996). Intonation allows the child to express a large number of communicative functions very early, long before he masters syntax. The way a child pronounces a sequence such as [pati papa oto] ("left dad car") indicates whether it is a joyful or desperate observation, or a question. Concerning this research area and that of developmental intonology, see for French examples Gabrielle Konopczynski, Konopczynski (1990).

It is usual to divide the languages into four categories: tonal languages (tone languages, typically Mandarin Chinese), stress languages (typically English), pitch-accent languages (typically Swedish and Japanese) and more recently boundary languages (typically French and the languages with distinctive lexical stress).

- In a language with lexical tones (still called *tonemes* to highlight their similarity with the phonemes), two syllables composed of the same phonemes will have two different meanings due to the difference of their lexical tone. In Mandarin Chinese — a classic example of tone languages — the syllable *ma* can have five different meanings depending on which of the four lexical tones is used: high tone, rising tone, falling-rising tone, falling tone, or no tone. Each of these tones is realized mostly by a particular pitch contour (F_0), which may be modified by the context. In certain languages, tones is not only a matter of F_0 contour. It also include a voice quality specification; for example, final glottal constriction; and in languages where they are defined solely by the F_0 contour, they nevertheless have, at the phonetic level, certain secondary features of duration, voice quality, and non-categorical modifications of the segments such as the timber of vowels and the articulation of consonants. For example, tone 3, In Chinese, has the longest realization and creaky voice may be associated with tone 3 production. The majority of the world's languages have tones.

- In a language with *lexical stress* like English, German, or Italian, each word has a lexical stress. Two words with the same sequence of phonemes may be distinguished by the position of the syllable that carries the primary lexical stress. In Russian (a free accent language), ***My**ka* ("tortoise") is opposed to *My**kà*** ("flour"). Lexical stress has various phonetic correlates that include, in different proportions depending on the language, fundamental frequency, duration, intensity, reduction of the timber of unaccented vowels (or complete change of the vowel color, such as in Russian), vowel harmony, and constraints on the distribution of phonemes. For example, three of the five Russian vowels /u, i, e, a, o/ found in stressed position show two levels of reduction in unstressed position : /e, a, o/ change their timber immediately before the stress, and after it. The opposition between /o/ and /a/ is lost in unstressed syllable. Not only the F_0 contour, but also the duration, the intensity and the realization of all the phonemes composing the word are generally modified by the position of the stressed syllable in a language with *lexical stress.* The stressed syllable does not receive a particular

F_0 contour through the lexicon, unlike in tone languages (like Chinese): its F_0 contour in stress languages is determined by phenomena of intonational nature. Inspired by the ideas of Mario Rossi, Rossi (1985) , we would say that the stressed syllable is a preferred site for anchoring *intonational morphemes* (but note that the word final syllable may also receive a continuation rise in English). Lexical stress is an abstract notion. It is an intrinsic feature of words and morphemes, stored in the mental lexicon, such as the tonemes in the so-called tone languages.

In addition to lexical stress there is *grammatical stress* — non-free stress, of morphological origin, or *demarcating stress* , not stored in the mental lexicon (Paul Garde, Larry Hyman). Garde (1968) Hyman (1977). In some languages, certain morphemes can influence the stress position in the word. For example in Italian the suffix -in- attracts the stress (con'tino: "little tale") while the morpheme -ic- pushes it back ('civico "civic").

- In Japanese and Swedish, the so-called languages with a melodic accent, called *pitch accent* languages, one of the syllables in Swedish or one of the *mora* in Japanese is marked by a pitch accent. The F_0 pattern of the word (that is anchored on the marked syllable) determines the F_0 contour of this syllable and the surrounding syllables in the word. The duration and the intensity are not profoundly modified by the position of the marked syllable in Japanese. In these two languages, the pitch accent domain is the word. (Note that in Japanese, some words are without accent).

- French is a ”boundary language”. In French, a language that has neither lexical tones nor lexical stress, the F_0 , timber duration variations are mainly anchored on the boundaries of words, syntagma and utterances. The last syllable is dominant, followed by the first syllable. The other syllables are uttered with the least strength. In Czech and Finnish, the ”stress” has also a demarcating function and always occurs on the first syllable of words. In Polish, as well, the stress is demarcating and is almost always on the penultimate syllable.

Stress, understood in a broad sense including both lexical and grammatical, and boundary create a *structure of dependence* between the syllables of the word, participating to give to the ”word” an acoustic reality. The most dominant syllable within a semantic unit or word tends to impose some of its features on the surrounding syllables. For example, in French, where the position of the stress is not distinctive, the final syllable of the word is in a dominant position without, however, always being perceived as prominent in the word. Some of the distinctive features of the stressed syllable tend to spread over the whole word (but it is not mandatory), such as nasality (*maman* /mamɑ̃/ pronounced [mɑ̃mɑ̃] "Mum" , or aperture (*était* /etɛ/ "was" pronounced [ete]; *phonologue* ”phonologist” pronounced [fɔnɔlɔg] (all vowels are semi-open) versus *phonologie* ”phonologie” pronounced [fonoloʒi] (all vowels are close or semi close). For children, *surtout* [syʁtu] "especially" is pronounced *sourtout*, [suʁtu], and *petit* /pəti/ ”small” pronounced [piti]). In some languages like Hungarian and Turkish, the same tendency of a feature to spread from a syllable to the whole word has been phonologized (i.e. made compulsory) in the form of, for example, a *vowel harmony*. That is to say, all the vowels (with some exceptions) of a word must share the same feature of frontness or backness of the tongue position, lip rounding or lip spreading. Vowel harmony participates in the perceptual integrity of the word.

This *structure of dependence* is essential for explaining the phenomena described by *historical phonetics* , which constitute the largest existing database on the long-term effect of word stress and word boundaries, on the one hand, and articulation of phonemes on the other. The most dominant phonemes in the word persisted while the least dominant phonemes are reduced merged with another surrounding phoneme or even disappear with time. In the passage from Latin to French, only the stressed syllable (in general the penultimate syllable in classical Latin) and the first syllable resisted wear in most frequently used words, the other ones tend to disappear: MUS*culum* > muscle > *moule* "muscle", CLA*ri*__ta__*tem* > *claret* "clarity" (for a review of different types of ”stress” to explain sound changes Vaissière (2001).

The *syllable* is also part of the *structure of dependence* . The stress and boundary structure of words determine the relationships of dependence between the phonemes constituting the syllable. For example, the consonant in coda depends on the preceding vowel generally more than the initial consonant depends on the following vowel. These degree of interdependence of the phonemes inside the syllable vary from language to language. The relation between a vowel and the following coda consonant within the rhyme at the end of the word are stronger in English (e.g. between /ɪ/ and /t/ in *sit*) than in French. In French, the word-final consonant tends to detach itself from the syllabic rhyme to which it belongs by ”enchaînement”. or by being released. In the case of ”enchaînement” the consonant at the end of a word is transferred onto the next word, if the next word starts with a vowel: *madame est …* [ma-da-me-], rather than [ma-dam-e-]. When the next word does start with a vowel, the consonant at the end of a sense-group tends to be strongly released to create what sounds like a next third syllable /madam/ > [ma-da-mə]. ”Enchaînement” and clear release are difficult to realise for natives of English, whose language has mostly closed syllables. English speakers, on the contrary, make a clear difference between the pronunciation of *an aim* [ən-e^{i}m] and *a name* [ə-ne^{i}m], a difference that is very difficult for a native French speaker to realize, whose language has mostly open syllables: he will pronounce ”an aim” as ”a name” (Delattre & Olsen (1969).

Insistence on a particular function of prosody varies according to the type of corpus on which the researcher worked: the reading of isolated sentences, the reading of text, or a conversation.

1. The analysis of isolated sentences confirms the known effect of *syntax, pragmatics and length of the constituents* . The reading of *isolated or even ambiguous sentences* highlights its demarcating function, which is related to syntax: the major boundary tends to move between the subject and the verb, *L'écolier / part à l'école* “The schoolboy is going to school.” (The slash indicates here the position of a boundary.) The position of the main boundary respects the linguistic surface structure. The analysis of the answers to questions of the type *Où part l'écolier ?* “Where is the schoolboy going?” reveals how the pragmatic division can significantly change the demarcation of syntactic origin because, in the end, pragmatics dominates in the utterance since it is semantic in nature: *l'écolier part/ à l'école* as an answer to the question “Where is the schoolboy is going to?.” The elements providing an answer to the question may also be focused (i.e. highlighted). The speaker tends to balance the length of the constituents, Grosjean (2011). *La reine d'Angleterre / visitera Paris demain.* ”The Queen of England / will visit Paris tomorrow.” versus *Le roi visitera / Paris demain.* ”The King of England / will visit Paris tomorrow.”

2. The reading of *texts* reveals the existence of a structure above the level of the sentence. Paragraphs possess a prosodic structure that indicates the beginning and end of paragraphs and characterizes the body of the paragraph, Lehiste (1975).

3. The study of a *life conversation* and *spontaneous speech* will illustrate the *discursive function of prosody in its broadest sense* . In this regard, prosody helps to distinguish information already shared between the speaker and the listener from new information and gives clues about the information that could possibly be called into question by the interlocutor. The new

> information in comparison to what has already been uttered is highlighted through prosodic processes. It also helps manage conversation *turn taking* and indicates that an assertion is final or waits for confirmation from the interlocutor. The way the word are spoken adds attitudinal nuances. For instance, the way a *no* is pronounced may indicated a categorical refusal or may suggest that refusal can be negotiated. A *no* can even sometimes mean a "yes". Theater voices highlight the identifying function of prosody, that is, the actors change their speaking style according to the roles they play. Its aesthetic function is clear in poetic expression.

All known languages, whatever their type of lexical stress, use intonational processes. Tone languages, in which the fundamental frequency is to some extent constrained by the lexical tone phenomenon, can use the duration and sound intensity of the syllables, as well as the F_0 register expansion, for intonational purposes. The presence of lexical tones by themselves is therefore not exclusive of intonational phenomena, and it is not possible to oppose *languages with lexical tones* and *languages with intonation*, as is sometimes done. Nevertheless, in tone languages such as Chinese or Vietnamese, the expression of modalities (interrogatives and statements) and attitudes is done by means of discourse particles as well as by prosody.

Some acoustic similarities can be observed in the F_0 contours in a number of languages, as illustrated in the figure below.

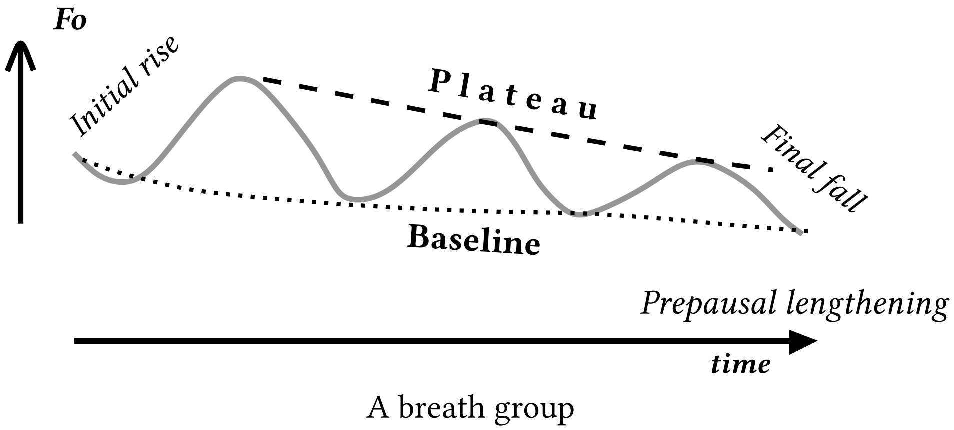


Figure 9.1: The typical F_0 contour of the utterance in various languages (According to Vaissière (1983)

Concerning similarities, the common tendencies are as follows:

1. The F_0 values vary between two lines, the baseline and the plateau, which limit the usual range of the speaker. The baseline is less variable than the plateau but can still be raised.

2. F_0 values and the sound intensity, as well as the precision of the articulatory gestures, tend to decline regularly over time.

3. Maximum F_0 and intensity values tend to be found in the first three syllables of the utterance, and the very first syllables of the utterance tend to have an increasingly higher F_0 and an increasingly greater intensity, corresponding to lexical or grammatical words. The minimum intensity value is at the end of the utterance, before the final pause of the utterance.

4. There is a tendency to periodically alternate the rises, or jumps, and falls of the F_0: a rise-fall pair tends to limit a unit of meaning in the broad sense, at least in non-tonal languages.

5. There is a tendency to lengthen the last syllable before a non final pause in an utterance and less regularly the final syllable of the utterance and to strengthen the first phoneme of the beginning of the sentence. The two tendencies are found at lower levels of prosodic constituents, such as the syntagma or the prosodic word. The overall shape and intensity of F_0 resemble the contour of the cries of babies, as well as the productions of certain monkeys; it seems to be physiologically determined. These common tendencies seem to have been used to characterize assertive utterances in the world's languages. For universal prosodic features, see Vaissière (1983).

Every language has its own preferred variety of this prototypical form to mark lexical stress and/or intonational morphemes. In French, it is the rising movement with a concomitant final lengthening at the end of a sense group, corresponding to a continuative morpheme, which perceptually stands out as a prototypical form. In Japanese, it is the downstep between two vowels of F_0 between two moras that becomes systematic in the realization of the melodic stress of the word, followed by a sudden upstep (jump) in one of the following syllables while in Danish the trough of F_0 followed by a rise marks the realization of the stressed syllable.

The physiological characteristics of the breath group (i.e. chunk of speech between two breathing pauses) appear to have motivated a certain number of mental associations related to boundary marking. A high or rising F_0 value and a strong intensity evoke the notion of beginning, that is to say, the beginning of a paragraph, of an utterance, of an intonational phase. A low or falling F $_0$, a weak intensity and a slow speech rate mark the end of a stretch of speech. A rising of the baseline or stopping of the F $_0$ declination in the course of an utterance without a breath pause simulates rebreathing and marks a boundary. An increase in the F $_0$ value signals the importance of what is said. Observations on various languages tend to confirm these general reflections, while showing in detail a great variety of achievements(For a review, Vaissière (1995).

The *frequency code* explains certain prosodic tendencies common to the most diverse languages. There is a biological association between a low F$_0$ and a large larynx (the low F$_0$ evokes thick vocal folds) and, conversely, between a high F$_0$ and a small larynx. The dominant male monkey emits lower sounds than the monkey that signals its submission with higher sounds, and the female monkey emits higher sounds when speaking to her newborn than to her older children. A low F$_0$ evokes maturity, dominance, and aggressiveness. The frequence code makes the hypothesis than these associations have been transposed in human speech. In languages, a low F$_0$ is a component of the intonemes used to mark orders and categorical affirmations, which evoke a sense of dominance. A high F$_0$, on the other hand, is an acoustic marker of uncertainty, doubt, questioning, an unfinished character of utterances, politeness and the desire to please, and a certain form of femininity. Thus, a well-attested behavior in monkeys is found as one of the ingredients of the complex game that constitutes intonation. Progress in the field of intonation studies is doubtless due to the multiplication of such insights, starting from the conviction that there is no mystery in the matter, while recognizing the complexity of the skein that the intonologist tries to untangle (n the frequency code, (Ohala (1994), Morton (2006).

[1].

A large number of similarities between languages appear in their use of prosodic parameters with some notable exceptions[2].

1. *Emotional* processes are strongly motivated by physiology and differ little between languages, at least in regards to the expression of primary emotions like joy or anger.

2. *The expression of attitudes* , however, often seems to make use of the same clues in a large number of languages: for example, an increase in the F$_0$ variation range of the whole utterance indicates a strong involvement of

[1]

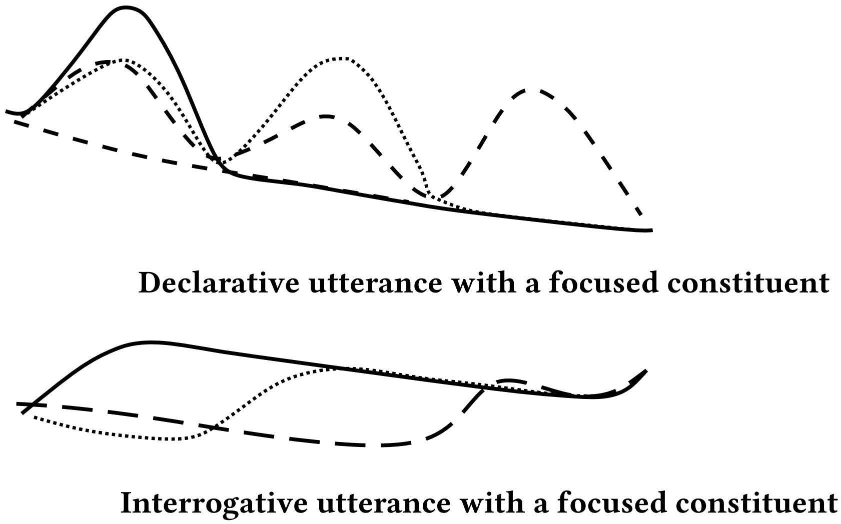
[2] See Annie Rialland's works on African languages, on intonation in African languages, see Downing & Rialland (2016)

the speaker. A particular F_0 variation on a single vowel (glissando) conveys an affective meaning. The increase in the intensity value or F_0 range, and the amplitude of the movements of articulators (greater closure for consonants and larger opening for vowels) simulates a greater respiratory, phonatory and articulatory effort, called the *code of effort* in Carlos Gussenhoven's terminology, Gussenhoven (2002). This increase is therefore interpreted by the listener as a sign of a greater involvement of the speaker, that is, the speaker makes more effort on the parts of the speech that he considers more important. The marking of *attitudes* is less directly motivated and the oral expression code of certain attitudes (such as irony) should be explained to second language learners.

3. Many languages are similar in the use of processes marking *focus in declarative and interrogative sentences*: the *focus* being characterized, both consensually and vaguely, as for example *what the interrogative structure concerns in a question, and what the assertion concerns in an affirmation.* Figure 9.2 illustrates the typical contours often attested in declarative and interrogative sentences in French, English or Japanese, with one focused word out of three.

It should also be noted that to a sentence or speech act corresponds a *whole prosodic pattern* , in which all constituents are linked. For example, a word may become perceptually prominent by its intrinsic acoustic properties, such as a higher F_0 and/or a longer duration and/or higher intensity (positive process) or by the de-stressing of the surrounding words, in particular the words that follow it (negative process). In a declarative utterance, the F_0 contour flattens after the realization of the focused word and remains in the low register. After the realization of the focused word, temporal changes such as the lengthening of lexically stressed syllables or the lengthening of the last syllable of words can replace the F_0 task to structure the remainder of the utterance (in the part with reduced F_0 range of variation), but such strategies of lengthening are not systematically observed. Furthermore, a listener is accurate at predicting the number of words of the upcoming endings in an utterance, basing himself solely on prosodic cues present at the sentence beginning, Grosjean (1983). While a declining F_0 curve and a final fall are observed in declarative sentences, the omission or decrease of the declination line starts from the sentence beginning (See Nina Thorsen for the Danish, Thorsen (1980)).

The acceleration of the speech rate reduces the perceptual obviousness of the prosodic structure and the marked number of prosodic constituent levels, and



. and drawn stroke to the case where the emphasis is placed on the last word.

Figure 9.2: The schematic representation of typical F_0 contours in declarative and interrogative sentences composed of three words, one of which is focused: solid line corresponds to the case where the emphasis is placed on the first word, dotted line to the case where the emphasis is placed on the second word, the second (), or the last word (drawn stroke)

consequently the prosodic structuring can become unrecognizable in very rapid speech. For instance, at a very fast rate, the F_0 flattens and there will be fewer cases of syllable lengthening. In such a situation, only the organization of the message into utterances can be transparent, mainly thanks to the pauses separating them, and the lower prosodic levels won't be marked. Another tendency that contributes to the dissociation of the syntactic structure from the pragmatic structure on the one hand, and from the prosodic structure on the other, is the tendency of prosodic constituents to be of equal size, as mentionned gy Maurice Grammont. The intonation units tend to be rhythmically balanced, i.e. to have the same number of syllables. For example, although the subject of the sentence normally carries the major continuation intoneme, as in *L'écolier / part à l'école*, a French speaker will more readily say *Jean part / à l'école*, in order to maintain a rhythmic balance between the two parts of the utterance. In this regard the resulting prosodic structuring does not reflect the syntactic-pragmatic structuring. The space between two stressed syllables (i.e. two stress groups) tends to equalize in English, and so does the length of the words. The duration of phonemes tend to shorten when the number of syllables in the word increases (Nooteboom (1972); Klatt (1973)).

Intonational choices of stylistic nature may have a negative impact on other components of prosody. Thus, in French, speakers (presenters, politicians or teachers) often emphasize the word- initial syllable (la **si**tuation du **pré**sident …) ”The president's situation”. The proliferation of emphatic stress aims to show the speaker's personal involvement in his discourse. If these initial boundary marks of the word help the listener to cut the discourse into words, by marking their beginning, the intrusion of this strong initial emphasis significantly alters the *traditional* French rhythm based on the recurrence of lengthened (final) syllables and rising F_0 in French. In many languages, grammatical words (including articles and auxiliaries, etc.) are weakly realized, such in French. Style may create exceptions to this general tendency. In that case, grammatical words may be emphasized (**la** situation **du** président …). Another difficulty is that speech accidents disturb rhythm and make its description difficult. This is the case of false starts, silent or filled hesitations (of the type *Papa euh vient* or the lengthening of the last syllable of the word of the type *Papaaaaaaa vient*) as well as that of stylistic choices: pauses of insistence (that is, the pause before a word with the purpose of drawing attention to it) and pauses between utterances. Politicians, once they have been elected, pause for longer periods and pause more frequently than during their election campaign, as shown by Danielle Duez's studies on French, Duez (1997).

The *similarity* of processes between languages also appears in the organization of the sentence into intonation groups Vaissière (1995). Figure 9.3 presents two typical examples of an utterance in French and English, two languages in which segmental and prosodic characteristics are generally recognized as opposite. Each sentence is composed of two intonational groups (breath group), each followed by a pause. The two intonation groups are separated by a non final pause. Each intonation group here includes two prosodic syntagma containing each two prosodic words.

The differences are the following:

1. The first difference between French and English concerns the *intoneme or morpheme of continuation*, at the end of the first (non final) intonation group, In both languages, it is realized by a rising F_0 contour on the last syllable, which is lengthened, and by a lengthening of the interval with the following intonation group. In French, the continuation rise is much more marked, and quasi-obligatory (see the arrow in Fig. 20). In English it is reduced and optional, called a *hook* by Pierre Delattre, Delattre (1963).

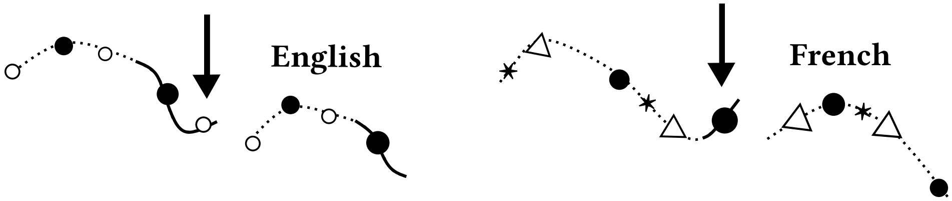


Figure 9.3: The schematic representation of the typical F_0 contour with the division of a breath group into two intonation groups, in English on the left, and in French on the right. The arrow indicates the continuation rise before the non final pause. Solid circles (•) correspond to lexically four stressed vowels in English and to the last vowel of the four words in French. Empty circles (O) represent unstressed syllables of a lexical word in English; Triangles (Δ) mark the beginning of a lexical word in French;   Asterisks (*) designs a grammatical word in both languages

2. The second remark concerns *the anchoring of the F_0 movements.* In English (where each lexical word has a lexically stressed syllable) the F_0 movements are anchored mainly on lexically stressed syllables and secondarily on the final syllable of words such as in the case of the continuation rise. In French (where there is no lexical stress), the F_0 movements are anchored on the two extreme syllables of the word: on the final syllable (essentially a major or minor continuation rise, or a peak) and on the first syllable, a non obligatory F_0 jump.

3. The third remark concerns the way how each language typically *regroups two words* into a single prosodic syntagma in a neutral declarative sentence. In English, the stressed syllable of the first word constituting the syntagma in a neutral declarative sentence has a rising movement or a high F_0 value. In addition, the stressed syllable of the second word has a falling movement, or a high value on that syllable followed by a fall on the following syllables. The first rise (or peak) on the stressed syllable of the first word the fall on the stressed syllable of the second word and the intermediate plateau form the so-called *hat pattern* (Shinji Maeda, Maeda (1976). The rise/plateau/fall shape is also described as H* H* L- in Tobi Silverman et al. (1992). In the French example illustrated here, the prosodic cue of the division of the syntagma into prosodic words is not provided through an F_0 movement but through the lengthening of the last syllable of lexical words Delattre (1966a) : the two regrouped words tend to have the F_0 pattern of a single word.

> In these two languages, the raising of the baseline of the F_0 contour (the so-called resetting of the base-line or F_0 upstep) usually appears between the two higher-level constituents in the sentence, for example, between two propositions or two intonation groups of the utterance. In Japanese, it is the raising of the baseline between two syntagma (and two intonation groups) that has been phonologized as a left boundary marker. The other processes such as final lengthening appear in Japanese in some speech styles (particularly before a non-final pause) but the continuation rise has not been attested in this language.
>
> All languages seem to use the same ingredients with the same demarcating function, such as final lengthening, continuation rise, return to the baseline, baseline resetting, etc., and the same elements for marking ”stress” but in a different or a very different manner and resolve in a different way the conflicts (for a review, Vaissière (1995).

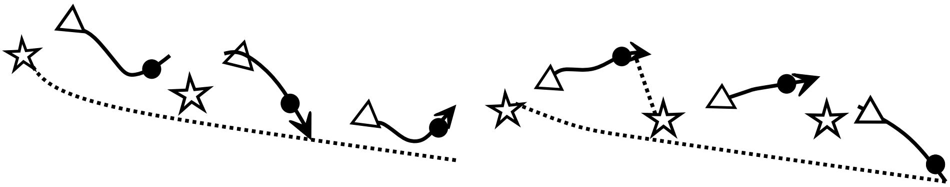
Language rhythm, as another prosodic component, is a very difficult notion to define. Rhythm is the perceived recurrence of a pattern over time and refers to perception.

1. *French* is often described as a language of *rising intonation*, in reference to the realizations of major and minor continuation rises at the end of many words situated at the end of sense groups (Pierre Delattre). What a French ear seems essentially to retain from the melody of an utterance is the repetition in time of the *continuation morpheme* at the end of the prosodic syntagma, realized by a melodic rise accompanied by a lengthening of the final rhyme. Vowels in the final position of sense groups and intonational groups dominate perceptually in French. It is the recurrence of lengthened vowels and rising intonations that actually define the rhythm in French.

2. In *English* the main perceptual rhythm unit is the *stress group*, starting with a stressed syllable, followed by a series of unstressed syllables. What strikes a Frenchman listening to English, is the energetic and quasi-regular recurrence of strongly stressed syllables, with a strong consonant onset and a F_0 fall on the stressed syllables, alternating with reduced syllables. For the French ear, this type of stress evokes the French emphatic stress (*l'accent d'insistance français*), which is often realized at the beginning of the word, hence the strange impression of insistence on every lexical word that the English language can give to an untrained ear. On the other way, French listeners are sometimes considered as “deaf” to stress contrasts and have difficulties to locate it Dupoux et al. (2001).

3. Conversely, in *Japanese*, the rhythm may appear somewhat monotone, inexpressive, due to alternating sequences of series of high followed by a series of low syllables, without dominant syllables. The rhythm appears also somewhat chaotic, because the lengthening of some vowels does not seem to be related to boundary phenomena, as expected by a native of French (the duration of vowels in Japanese depends primarily on their phonological duration) and is therefore not correlated with the melodic movements — unlike French — and on the realization of a lexical stress unlike English (for comparison between French and English rhythm, Vaissière (1991).

Let us give some examples of the use of prosody in French. Figure 9.4 illustrates the general tendencies in French.



⊠ ⋆ : grammatical word;

Figure 9.4: The schematic representation of the classical division of the French sentence into two breath groups: each breath group is divided here into three prosodic words. Δ : beginning of a lexical word; • : final syllable of a lexical word

The following remarks can be made:

1. the non final breath group ends with a sharp continuation rise on the last syllable, after a return to the baseline,
2. the final breath group ends with a final fall over several syllables starting from the last syllable of the penultimate lexical word (same adjustment is necessary when the last lexical word is monosyllabic to insure a F_0 descent on several syllables)
3. the penultimate lexical word ends with an high F_0 value or with a rise independently from its syntactic relationship with the following word, a necessary condition before the realisation of the final fall,
4. the lexical words tend to start with a F_0 jump; they may end with a rising, flat or a falling Fo pattern except for the rising F_0 pattern for the word before the non-final pause and for the penultimate word, and falling F_0 pattern for the final word in the sentence;

5. there is a resetting of the baseline during the non final pause

Figure 9.5 illustrates the F_0 and duration differences between three utterances that are homophones in terms of their constituent phonemes /sɛtɔmɛtenɔʁmemɑbɛt/ but correspond to three different sentences with a different syntactic structure. The interpretation of the sequence of phonemes is guided by the parameters of duration and F_0 defining major and minor boundaries:

“This man is enormously dumb.” /sɛtɔm||ɛtenɔʁmemɑbɛt/ (top),

“This man |is huge and bothers me.” /sɛtɔmɛtenɔʁm||emɑbɛt/ (mid)

or “This man and Tenor like me as a beast.” /sɛtɔm|ɛtenɔʁ||mem|ɑbɛt/ (bottom).

This highly oversimplified example has the advantage of allowing a direct comparison of observations. It is clear that the most rising syllable of the utterance corresponds to the major intonational boundary, realized on the last rhyme of the word (a syllable can be divided into onset and rhyme). The lengthening of the syllable that takes the rise reinforces the rise perceptually, and the continuation rise tends to extend on the sonorant coda.

The basic demarcation principle in French is very simple: within the utterance, the longer the last syllable of a word, the more the F_0 superimposed to this syllable rises and the more the intonational boundary is perceptually evident. A falling contour at the end of a word indicates, on the contrary, a dependence of this word on the next one: for example, the falling contour corresponding to an attributive adjective before the noun it completes.

Note that this dependence does not correspond strictly to the case highlighted by the syntactitians. More generally, the intonational boundaries do not mechanically reflect the syntactic structure, as mentioned before Grosjean (2011). The speaker is free, for the same utterance, to group together a sequence of several words into a single pattern corresponding to a single word Grammont (1920). For example, speech expressions manipulated by a computer show that the relative duration of the first syllable is enough to distinguish between *bordures* [bɔʁdy:: ʁ] [3]. “borders”, and *bords durs* [bɔ:ʁdy :: ʁ] “hard edges”, *Jean-Pierre et Jacques* [ʒɑpjɛ:ʁ e ʒa:k] and *Jean, Pierre et Jacques* [ʒɑ: pjɛ:ʁ eʒa:k], with no need to modify F_0: lengthening is powerful enough to indicate if the sense-group is composed by on word (*bordures*), two words (*bords durs*, *Jean-Pierre et Jacques*) or three words (*Jean, Pierre et Jacques*). The demarcative function is insured by both F_0 and duration. A single F_0 pattern may be subdivided into two parts by duration only (as exemplified just above); a single rhythmic unit (sequence of syllables lengthened at the end) may be decomposed into two parts but an F_0 movement.

[3] Note that the two points : indicate the degree of lengthening, and the repetition of this same symbol (: :) indicates a high degree of lengthening.

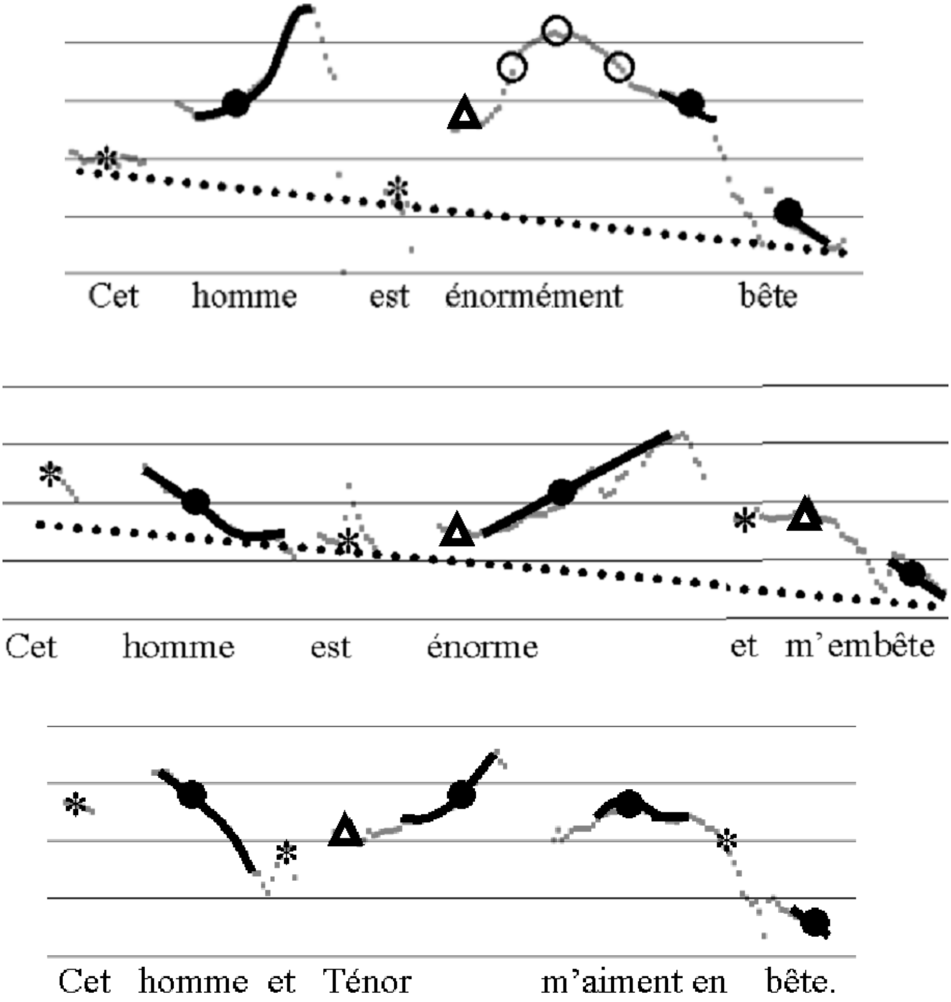


Figure 9.5: Top: *Cet homme est énormément bête.* “This man is very dumb.”; Middle: *Cet homme est énorme et m'embête.* “This man is huge and bothers me.” Bottom: *Cet homme et Ténor m'aiment en bête.* “This man and Tenor like me as a beast.”

Similarly, in English, the only relative duration of the second syllable *fee* is sufficient to distinguish between *coffee cake* and *honey*, and *coffee, cake and honey*. See Ilse Lehiste, Lehiste (1960) for the study of internal open juncture in English).

In most cases, the F_0 variations reinforce the division marked by duration contrasts.

Figure 9.6 (top) illustrates the most attested tendency of contrast between interrogative and declarative sentences in French, as well as unfinished sentences. The assertive sentence ends with a final fall. The unfinished sentence (dash-dotted line) and the interrogative sentence (dotted line) are opposed to the declarative by the presence of a final F_0 rise that expresses their unfinished character. Interrogative and unfinished sentences are differentiated typically by the fact that in the interrogative utterance there is no tendency for F_0 to decline and F_0 reaches the baseline before the final rise in the case of the unfinished sentence. The interrogatives are also pronounced faster than their declarative counterparts.

The same figure illustrates the F_0 contour superimposed to ”Marie vient à Paris demain”, as a question (top), as an unfinished sentence (middle) and as an assertion (bottom°.

Further readings Kawaguchi et al. (2006)

What to conclude?

The action of performance factors and the multiplicity of functions of intonation call for great care in the development of a *grammar of prosody*, a theory or a model. The prosody of a spontaneous utterance is shaped by too many factors, often unpredictable. Furthermore these different factors sometimes inextricably modify the same prosodic parameters such as F_0, intensity and duration.

There is at present no automated reliable system for the recognition of the prosodic structure, emotions or attitudes, as there are systems for speech recognition.

Trying to use the same labels for describing the prosody of a large number of languages is challenging and allows a better understanding of prosodic similarities and differences between languages. We recommand the reading of Hirst & Di Cristo (1998), Sun-Ah (2005), Sun-Ah (2014).

Prosodic studies have been at the forefront of professional meetings and journals.

We recommend the reading of the proceedings of the biennial meeting of the Speech Prosody Special Interest Group (SProSIG) of the International Speech Communication Association (ISCA), covers all aspects of prosody in spoken language.

New instrumentation permitting the visualization of the entire vocal tract in motion including vocal folds, complex computer programs (such as prosodic

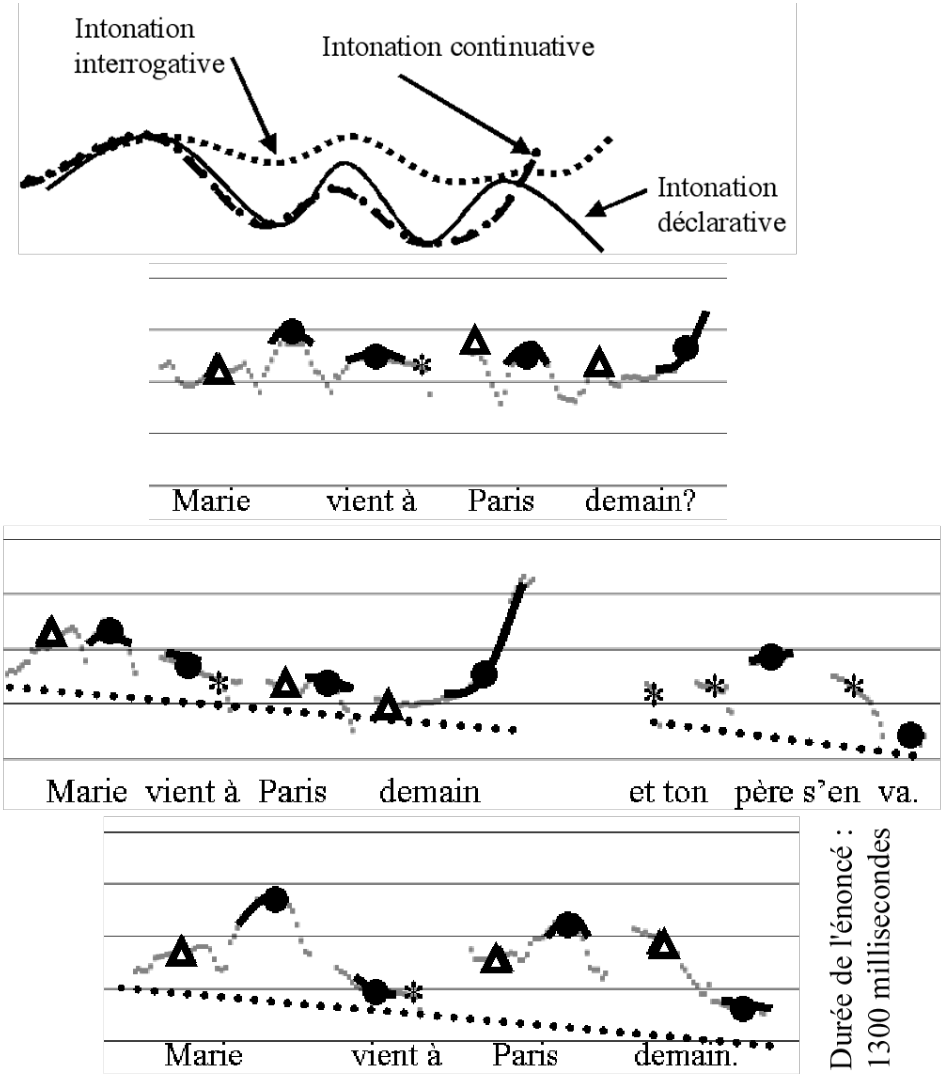


Figure 9.6: Top: Typical forms for declarative sentences (solid line), for non final intonational group (dash-dotted line) and for yes-no questions (dotted line). Middle top: example of interrogative sentence: *Marie vient à Paris demain?* "Mary is coming to Paris tomorrow." Middle: example of an assertive or declarative sentence: *Marie vient à Paris demain.* "Mary is coming to Paris tomorrow.". Bottom: example of a unfinished sentence followed by a non final pause: *Marie vient à Paris demain et ton père s'en va.* "Mary is coming to Paris tomorrow and your father is going away."

transpositions from one language to another), Advances in cerebral imaging now allow advanced studies on the interaction between prosody and the articulation of segments and on cerebral structures involved in prosodic processes.

10 Conclusion

This book has achieved its objectives if it has managed to show the extent of the results obtained by phonetics and speech sciences as well as the extent of the perspectives they broaden. Examples include the various scientifically established results on speech perception, to eventually be tested in cerebral imagery, with an ontogenetic and phylogenetic perspective, including the transposition of proven methods for the analysis and evaluation of the voice, normal and pathological speech, to methods of learning (attitudinal or behavioral) and aesthetic aspects for foreign language learners. The systematic use of articulatory synthesis programs in the future will finally allow the consideration of an integrated definition (articulatory, acoustic and perceptual) of the distinctive features proposed by Roman Jakobson. Knowledge of the exact function of each organ in the speech act studied by articulatory phonetics is useful for ENT physicians to predict the consequences of their surgical actions on speech and to explain it to their patients. Research on voice and speech disorders is the source of knowledge in this field. In addition, the advances in perceptual-acoustic studies have a direct impact on the development of cochlear implants. Likewise, the study of speech learning by cochlear implants allows the evaluation of the role of perception on first language acquisition.

Phonetic knowledge is within everyone's reach. Pedagogues, teachers, ENT and speech therapists, if they possessed a better background in phonetics than currently provided through formal education, could have a better understanding of the difficulties experienced in the practice of their profession and could then devise more suitable solutions.

What phonetics explores with the methods of the so-called hard sciences is the eminently human reality of speech, in the variety of its manifestations. The part accorded to prosody in this study was intended to highlight this component, which is specific to speech in comparison with the written text. The phenomena of orality are of great complexity, whether they are considered from an acoustic, physiological or perceptual point of view. This imposes on the professional phonetician a long and necessarily fragmentary training. The interpretation of the results furnished by experimental methods of increasing complexity imposes an in-depth specialization. Information technology has nevertheless made this training more readily available.

Only a multidisciplinary team, and not an individual, can bring cumulative progress to phonetic knowledge. In fundamental research in phonetics, the contribution of everyone (the phonologist, the psychologist, and the engineer, along with other specialists) is irreplaceable and also helps in the renewal of research questions. New technologies and their applications will continue to guide the reflections of phoneticians. As a meeting place between disciplines, phonetics today retains its status as a pilot science within the field of language sciences.

xxx To increase knowledge in this area, this review offers several recommendations, namely: that description of less studied populations should be prioritised; that the potential for automated methods for detecting creaky voice to expand the scope, replicability, and comparability of studies should be explored; that the relative merits of different prevalence formulae used to date should be considered; and, finally, that findings should be presented in such a way that is maximally insightful. xxx

References

Marchal, Alain & Christian Cavé (eds.). 2009. *L'imagerie médicale pour l'etude de la parole*. [On new techniques, their benefits and limitations]. Paris: Lavoisier-Hermès.

Aboitiz, Francisco. 2017. *A brain for speech: A view from evolutionary neuroanatomy*. Springer.

Anderson, Stephen R., Patricia A. Keating, Marie K. Huffman & Rena A. Krakow. 2014. *Nasals, nasalization, and the velum*. Vol. 5. Elsevier.

Association, International Phonetic. 1999. *Handbook of the international phonetic association: A guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

Baer, Thomas, John C. Gore, L. Carol Gracco & Patrick W. Nye. 1991. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *The Journal of the Acoustical Society of America* 90(2). 799–828.

Ball, J. B. & C. Code (eds.). 1997. *Instrumental clinical phonetics*. London: Whurr Publishers.

Ball, Martin J & Orla Lowry. 2008. *Methods in clinical phonetics*. John Wiley & Sons.

Ball, Martin J. 2021. *Manual of clinical phonetics*. Routledge.

Ball, Martin J., Michael R. Perkins, Nicole Müller & Sara Howard. 2008. *The handbook of clinical linguistics*. Wiley Online Library.

Bauman-Waengler, J. 2012. *Articulatory and phonological impairment: A clinical focus*. 4th ed. London: Pearson.

Beddor, Patrice Speeter. 1993. The perception of nasal vowels. In *Nasals, nasalization, and the velum*, 171–196. Elsevier.

Benguerel, André-Pierre & Helen A. Cowan. 1974. Coarticulation of upper lip protrusion in french. *Phonetica* 30(1). 41–55.

Benveniste, Emile. 1953. Animal communication and human language: The language of the bees. *Diogenes* 1(1). 1–7.

Best, Catherine T., Gerald W. McRoberts & Nomathemba M. Sithole. 1988. Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of experimental psychology: Human perception and performance* 14(3). 345.

Bhaskararao, Peri & Peter Ladefoged. 1991. Two types of voiceless nasals. *Journal of the International Phonetic Association* 21(2). 80–88.

Birkholz, Peter, Dietmar Jackèl & Bernd J. Kroger. 2006. Construction and control of a three-dimensional vocal tract model. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, I–I.

Boë, Louis-Jean. 1999. Modelling the growth of the vocal tract vowel spaces of newly-born infants and adults: Consequences for ontogenesis and phylogenesis. In *Proceedings of the international congress of phonetic sciences*, vol. 3, 1–25.

Boersma, Paul. 2011. Praat: Doing phonetics by computer. *http://www. praat. org/*. Computer program.

Bolinger, Dwight. 1972. Accent is predictable (if you're a mind-reader). *Language*. 633–644.

Bolinger, Dwight. 1982. Intonation and its parts. *Language*. 505–533.

Bothorel, A., P. Simon, F. Wioland & J. P. Zerling. 1986. Cinéradiographie des voyelles et des consonnes du français (cineradiography of lade in french). *Trav. de l'Inst. de Phonétique de Strasbourg*.

Boysson-Bardies, Bénédicte de. 2009. How language comes to children. *Revista Brasileira de Psicanálise* 43(1). 97–103.

Broca, Paul. 1861. Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bulletin et Memoires de la Societe anatomique de Paris* 6. 330–357.

Brown, Gillian, Karen Currie & Joanne Kenworthy. 2015. *Questions of intonation*. Routledge.

Carlson, Rolf, Gunnar Fant & Björn Granström. 1974. Two-formant models, pitch, and vowel perception. *Acta Acustica united with Acustica* 31(6). 360–362.

Carlson, Rolf, Björn Granström & Gunnar Fant. 1970. Some studies concerning perception of isolated vowels. *Speech Transmission Laboratory Quarterly Progress and Status Report* 11(2-3). 19–35.

Chambers, Jack K. & Natalie Schilling. 2018. *The handbook of language variation and change*. John Wiley & Sons.

Cheour, Marie, Rita Ceponiene, Anne Lehtokoski, Aavo Luuk, Jüri Allik, Kimmo Alho & Risto Näätänen. 1998. Development of language-specific phoneme representations in the infant brain. *Nature neuroscience* 1(5). 351–353.

Chiba, T. & M. Kajiyama. 1941. *The vowel: Its nature and structure*. Kaiseikan, Tokyo.

Chistovich, Ludmilla A. & Valentina V. Lublinskaya. 1979. The 'center of gravity'effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing research* 1(3). 185–195.

Chomsky, Noam & Morris Halle. 1968. The sound pattern of english.

Choukri, Khalid, Valerie Mapelli & Jeffrey Allen. 1999. New developments within the european language resources association (elra). In *Eurospeech*.

Clements, George N. 2003. Feature economy in sound systems. *Phonology* 20(3). 287–333.

Clopper, Cynthia G. & David B. Pisoni. 2004. Some acoustic cues for the perceptual categorization of american english regional dialects. *journal of phonetics* 32(1). 111–140.

Cohen, Antoine & Sibout G. Nooteboom. 1975. *Structure and process in speech perception*. Springer.

Collier, Rene & J. 't Hart. 1975. The role of intonation in speech perception. In *Structure and process in speech perception*, 107–123. Springer.

Coulthard, Malcolm, Alison Johnson & David Wright. 2016. *An introduction to forensic linguistics: Language in evidence*. Routledge.

Coustenoble, Hélène Nathalie & Lilias Eveline Armstrong. 1934. *Studies in french intonation*. W. Heffer & sons, Limited.

Damico, Jack S., Nicole Müller & Martin John Ball. 2010. *The handbook of language and speech disorders*. Wiley Online Library.

Dart, Sarah N. 1998. Comparing french and english coronal consonant articulation. *Journal of phonetics* 26(1). 71–94.

Darwin's, Charles. 1859. On the origin of species. *published on* 24.

De Saussure, Ferdinand. 1957. Cours de linguistique générale (1908-1909). *Cahiers Ferdinand de Saussure* (15). 3–103.

De Saussure, Ferdinand. 1959. *Course in general linguistics, trans. wade baskin*. New York.

de Boysson-Bardies, Bénédicte. 1996. *Comment la parole vient aux enfants*. Odile Jacob.

Delattre, Pierre. 1951. *Principes de phonétique française: À l'usage des étudiants anglo-américains*. École Française D'Été, Middlebury College.

Delattre, Pierre. 1963. Comparing the prosodic features of english, german, spanish and french.

Delattre, Pierre. 1966a. Accent de mot et accent de groupe. In *Studies in french and comparative phonetics*, 69–72. De Gruyter.

Delattre, Pierre. 1966b. Les attributs acoustiques de la nasalité vocalique et consonantique. In *Studies in french and comparative phonetics*, 257–263. De Gruyter.

Delattre, Pierre. 1966c. Les dix intonations de base du français. *French review*. 1–14.

Delattre, Pierre & Donald C Freeman. 1968. A dialect study of american r's by x-ray motion picture.

Delattre, Pierre, Alvin M. Liberman, Franklin S. Cooper & Louis J. Gerstman. 1952. An experimental study of the acoustic determinants of vowel color: Observations on one-and two-formant vowels synthesized from spectrographic patterns. *Word* 8(3). 195–210.

Delattre, Pierre & Carroll Olsen. 1969. Syllabic features and phonic impression in english, german, french and spanish. *Lingua* 22. 160–175.

Delattre, Pierre C., Alvin M. Liberman & Franklin S. Cooper. 1955. Acoustic loci and transitional cues for consonants. *The journal of the acoustical society of America* 27(4). 769–773.

de Mareüil, Philippe Boula. 2010. *D'où viennent les accents régionaux?* Le Pommier.

Denes, Peter B., Peter Denes & Elliot Pinson. 1993. *The speech chain.* Macmillan.

Derwing, Tracey M. & Murray J. Munro. 2015. *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research.* Amsterdam: John Benjamins.

Detey, Sylvain & Isabelle Racine. 2015. The interphonology of contemporary french (IPFC): An international corpus-based L2 phonology research programme. In.

Disner, Sandra Ferrari. 1983. *Vowel quality: The relation between universal and language specific factors.* Vol. 58. Los Angeles: Phonetics Laboratory, Department of Linguistics, UCLA.

Doughthy, C. J. & M. H. Lonh (eds.). 2003. *The handbook of second language acquisition.* Oxford/New Jersey: Wiley-Blackwell.

Downing, Laura J. & Annie Rialland. 2016. *Intonation in african tone languages.* Vol. 24. Walter de Gruyter.

Duez, Danielle. 1997. Acoustic markers of political power. *Journal of Psycholinguistic Research* 26(6). 641–654.

Dupoux, Emmanuel, Sharon Peperkamp & Núria Sebastián-Gallés. 2001. A robust method to study stress "deafness". *The Journal of the Acoustical Society of America* 110(3). 1606–1618.

Durand, Jacques, Bernard Laks & Chantal Lyche. 2005. Un corpus numérisé pour la phonologie du français. In G. Williams (ed.), *La linguistique de corpus, pp*, 205–217. Rennes: PUR.

Durand, Jacques, Bernard Laks & Chantal Lyche. 2014. French phonology from a corpus perspective. In *The oxford handbook of corpus phonology.*

Durand, Marguerite. 1930. *Étude sur les phonèmes postérieurs dans une articulation parisienne.* H. Didier.

Dutoit, Thierry. 1997. *An introduction to text-to-speech synthesis*. Vol. 3. Springer Science & Business Media.

Esling, John, Scott Moisik, Allison Benner & Lise Crevier-Buchman. 2019. *Voice quality the laryngeal articulator model*. Cambridge University Press.

Fant, Gunnar. 1960a. *Acoustic theory of speech production*. Walter de Gruyter.

Fant, Gunnar. 1960b. *On the accoustics of speech*. förf.

Fant, Gunnar. 1973. Stops in CV-syllables. *Speech sounds and features*. 110–139.

Fant, Gunnar. 1975. Vocal-tract area and length perturbations. *STL-QPSR* 4(1975). 1–14.

Fant, Gunnar. 2006. *Speech acoustics and phonetics: Selected writings*. [Advanced level on speech acoustics]. Dordrecht/Boston/London: Kluwer Academic Publishers.

Fant, Gunnar & Stefan Pauli. 1974. Vocal tract cavity-mode relations viewed by spatial energy distributions. *The Journal of the Acoustical Society of America* 55(2). 384–384.

Fant, Gunnar & Arne Risberg. 1963. Auditory matching of vowels with two formant synthetic sounds. *STL-Quarterly Progress Status Report*. 7–11.

Fletcher, Samuel G., Martin J. McCutcheon & Matthew B. Wolf. 1975. Dynamic palatometry. *Journal of Speech and Hearing Research* 18(4). 812–819.

Fónagy, Ivan. 1982. *Situation et signification*. John Benjamins Publishing.

Fónagy, Ivan. 1983. *La vive voix: Essais de psycho-phonétique*. Vol. 20. Payot.

Fónagy, Ivan. 1989. On status and functions of intonation. *Acta Linguistica Hungarica* 39(1/4). 53–92.

Fónagy, Ivan. 1990. The chances of vocal characterology. *Acta Linguistica Hungarica* 40(3/4). 285–313.

Fónagy, Ivan, Eva Bérard & Judith Fónagy. 1983. Clichés mélodiques.

Fry, Dennis. B. (ed.). 1976. *Acoustic phonetics: A course of basic readings*. Cambridge: Cambridge University Press.

Fujimura, Osamu. 1962. Analysis of nasal consonants. *The Journal of the Acoustical Society of America* 34(12). 1865–1875.

Fujimura, Osamu, Shigeru Kiritani & Haruhisa Ishida. 1973. Computer controlled radiography for observation of movements of articulatory and other human organs. *Computers in Biology and Medicine* 3(4). 371–384.

Garde, Paul. 1968. *L'accent*. Vol. 5. Presses universitaires de France.

Gay, Thomas, Björn Lindblom & James Lubker. 1981. Production of bite-block vowels: Acoustic equivalence by selective compensation. *The Journal of the Acoustical Society of America* 69(3). 802–810.

Gick, Bryan, Ian Wilson, Karsten Koch & Clare Cook. 2004. Language-specific articulatory settings: Evidence from inter-utterance rest position. *Phonetica* 61(4). 220–233.

Gordon, Matthew & Peter Ladefoged. 2001. Phonation types: A cross-linguistic overview. *Journal of phonetics* 29(4). 383–406.

Grammont, Maurice. 1920. *Traité pratique de prononciation française*. Delagrave.

Grosjean, François. 1983. How long is the sentence? prediction and prosody in the on-line processing of language.

Grosjean, François. 2011. Linguistic structures and performance structures: Studies in pause distribution. In *Temporal variables in speech*, 91–106. De Gruyter Mouton.

Guenther, F. 2016. *Neural control of speech*. Cambridge/Massachusetts: The MIT Press.

Gussenhoven, Carlos. 2002. Intonation and interpretation: Phonetics and phonology. In *Speech prosody 2002, international conference*.

Handel, Stephen. 1993. *Listening: An introduction to the perception of auditory events*. Cambridge: The MIT Press.

Harrington, J. & M. Tabain (eds.). 2006. *Speech production: Models, phonetic processes, and techniques*. New York: Psychology Press.

Harris, Katherine Safford. 1954. Cues for the identification of the fricatives of american english. *The Journal of the Acoustical Society of America* 26(5). 952–952.

Hart, Johan't, René Collier & Antonie Cohen. 2006. *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge University Press.

Hayward, Katrina. 2014. *Experimental phonetics: An introduction*. Routledge.

Heeman, Peter A. & James Allen. 1999. Speech repains, intonational phrases, and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics* 25(4). 527–572.

Hirst, Daniel & Albert Di Cristo. 1998. *Intonation systems: A survey of twenty languages*. Cambridge University Press Cambridge.

Holmes, John & Wendy Holmes. 2002. *Speech synthesis and recognition*. CRC press.

Honeybone, P. & J. Salmons (eds.). 2015. *The Oxford handbook of historical phonology*. Oxford: OUP.

House, Arthur S. & Grant Fairbanks. 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America* 25(1). 105–113.

House, Arthur S. & Edward P. Neuburg. 1977. Toward automatic identification of the language of an utterance. i. preliminary methodological considerations. *The Journal of the Acoustical Society of America* 62(3). 708–713.

Hyman, Larry. 1977. On the nature of linguistic stress. *Studies in stress and accent* 4. 37–82.

Ishizaka, Kenzo & James L. Flanagan. 1972. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell system technical journal* 51(6). 1233–1268.

Jakobson, Roman, Gunnar Fant & Morris Halle. 1952. *Preliminaries to speech analysis.* Cambridge: The MIT Press.

Jakobson, Roman, Claude Lévi-Strauss & John Mepham. 1978. *Six lectures on sound and meaning.* Harvester Press Brighton.

Jessen, M. 2008. *Forensic phonetics.* Oxford/New Jersey: Wiley-Blackwell.

Jessen, Michael. 2002. *Forensic voice identification.* Cambridge/Massachusetts: Academic Press.

Johnson, Keith. 2012. *Acoustic and auditory phonetics.* 3rd edition. (1st ed. in 1997) [Basics of acoustic phonetics]. Oxford/New Jersey: Wiley-Blackwell.

Johnson, Keith, Peter Ladefoged & Mona Lindau. 1993. Individual differences in vowel production. *The Journal of the Acoustical Society of America* 94(2). 701–714.

Jongman, Allard, Sheila E. Blumstein & Aditi Lahiri. 1985. Acoustic properties for dental and alveolar stop consonants: A cross-language study. *Journal of Phonetics* 13(2). 235–251.

Joseph, B. & R. D. Janda (eds.). 2003. *The handbook of historical linguistics.* Oxford/New Jersey: Wiley-Blackwell.

Karcevski, Serge. 1931. Sur la phonologie de la phrase. *Travaux du Cercle linguistique de Prague* (4). 188–227.

Kawaguchi, Yuji, Iván Fónagy & Tsunekazu Moriguchi. 2006. *Prosody and syntax: Cross-linguistic perspectives.* Vol. 3. Amsterdam: John Benjamins Publishing.

Kent, R. D. & Ch Read. 1992. *The acoustic analysis of speech.* 2nd ed. (1st ed. in 1992) [Intermediate level]. London/San Diego: Whurr Publishers – Singular Publishing.

Kiritani, Shigeru, Kenji Itoh & Osamu Fujimura. 1975. Tongue-pellet tracking by a computer-controlled x-ray microbeam system. *The Journal of the Acoustical Society of America* 57(6). 1516–1520.

Klatt, Dennis H. 1973. Vowel duration as a function of the syllabic structure of a word. *The Journal of the Acoustical Society of America* 54(1). 312–313.

Klatt, Dennis H. 1980. Software for a cascade/parallel formant synthesizer. *the Journal of the Acoustical Society of America* 67(3). 971–995.

Kohler, Klaus J. 1990. Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In *Speech production and speech modelling*, 69–92. Springer.

Konopczynski, Gabrielle. 1990. *Le langage émergent: Aspects vocaux et mélodiques*. Vol. 60. Buske Verlag.

Krämer, Martin. 2008. *Vowel harmony and correspondence theory*. De Gruyter Mouton.

Kreiman, J. & D. Sidtis. 2013. *Foundations of voice studies*. Oxford/New Jersey: Wiley-Blackwell.

Kreiman, Jody & Bruce R. Gerratt. 2003. Jitter, shimmer, and noise in pathological voice quality perception. In *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*.

Kröger, Bernd J. & Peter Birkholz. 2009. Articulatory synthesis of speech and singing: State of the art and suggestions for future research. *Multimodal Signals: Cognitive and Algorithmic Issues*. 306–319.

Kuhl, Patricia K & Andrew N Meltzoff. 1982. The bimodal perception of speech in infancy. *Science* 218(4577). 1138–1141.

Kuhl, Patricia K. 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & psychophysics* 50(2). 93–107.

Labov, William. 1972. *Sociolinguistic patterns*. University of Pennsylvania press.

Ladefoged, Peter. 1967. *Three areas of experimental phonetics*. Oxford University Press.

Ladefoged, Peter. 1990. The revised international phonetic alphabet. *Language* 66(3). 550–552.

Ladefoged, Peter. 1996. *Elements of acoustic phonetics*. University of Chicago Press.

Ladefoged, Peter. 2003. *Phonetic data analysis. an introduction to fieldwork and instrumental techniques*. Oxford, Blackwell Publishing.

Ladefoged, Peter. 2005. *Vowels and consonants*. Vol. 1. Wiley-Blackwell.

Ladefoged, Peter & Donald Eric Broadbent. 1957. Information conveyed by vowels. *The Journal of the acoustical society of America* 29(1). 98–104.

Ladefoged, Peter & Keith Johnson. 2014. 4 cardinal vowels. *Pronunciation and Phonetics: A Practical Guide for English Language Teachers*. 20.

Ladefoged, Peter & Ian Maddieson. 1996. *The sounds of the world's languages*. [An important classic]. Massachusetts/Oxford: Blackwell.

Ladefoged, Peter & Ian Maddieson. 1998. The sounds of the world's languages. *Language* 74(2). 374–376.

Lass, Norman. 2012. *Contemporary issues in experimental phonetics*. Elsevier.

Lass, Norman J. (ed.). 1996. *Principles of experimental phonetics*. Missouri: Mosby.

Laver, John. 1978. The concept of articulatory settings: An historical survey. *Historiographia Linguistica* 5(1-2). 1–14.

Lee, Chin-Hui, Frank K. Soong & Kuldip K. Paliwal. 2012. *Automatic speech and speaker recognition: Advanced topics*. Vol. 355. Springer Science & Business Media.

Lehiste, Ilse. 1960. An acoustic–phonetic study of internal open juncture. *Phonetica* 5(s1). 5–54.

Lehiste, Ilse (ed.). 1967. *Readings in acoustic phonetics*. Cambridge/Massachusetts: The MIT Press.

Lehiste, Ilse. 1970. *Suprasegmentals*. Cambridge/Massachusetts: The MIT Press.

Lehiste, Ilse. 1975. The phonetic structure of paragraphs. In *Structure and process in speech perception*, 195–206. Springer.

Lehiste, Ilse & Gordon E. Peterson. 1959. Vowel amplitude and phonemic stress in american english. *The Journal of the Acoustical Society of America* 31(4). 428–435.

Léon, Pierre R. 1993. *Précis de phonostylistique: Parole et expressivité*. [On prosody]. Paris: Nathan Université.

Léon, Pierre. 2009. A new look at phonostylistics. *La linguistique* 45(1). 159–170.

Levelt, Willem J. M. 1989. *Speaking: From intention to articulation*. A more advanced presentation of the speech chain. Cambridge/Massachusetts: The MIT Press.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler & M. Studdert-Kennedy. 1967. Perception of the speech code. *Psychological Review* 74(6). [On categorical perception and the motor theory], 431–461.

Liberman, Alvin M. & Ignatius G. Mattingly. 1985. The motor theory of speech perception revised. *Cognition* 21(1). 1–36.

Liberman, Alvin Meyer. 1996. *Speech: A special code*. MIT press.

Liberman, Mark & Christopher Cieri. 1998. The creation, distribution and use of linguistic data: The case of the linguistic data consortium. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, 159–164.

Liénard, Jean-Sylvain. 1977. *Les processus de la communication parlée: Introduction à l'analyse et la synthèse de la parole*. Masson.

Lindblom, Björn. 1986. Phonetic universals in vowel systems. *Experimental phonology*. 13–44.

Lindblom, Björn & Ian Maddieson. 1988. Phonetic universals in consonant systems. *Language, speech and mind* 6278.

Lisker, Leigh. 1986. "voicing" in english: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and speech* 29(1). 3–11.

Lisker, Leigh & Arthur S. Abramson. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20(3). 384–422.

MacNeilage, P. (ed.). 1983. *The production of speech*. Berlin/Heidelberg/New York: Springer-Verlag.

Maddieson, I. 1984. *Patterns of sounds*. Cambridge: Cambridge University Press. [On the sounds of the world's languages].

Maddieson, Ian. 1981. *UPSID: UCLA Phonological Segment Inventory Database*. Los Angeles.

Maddieson, Ian. 2013. Voicing in plosives and fricatives. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/4.

Maeda, Shinji. 1976. *A characterization of American English intonation*. MIT. (Doctoral dissertation). One of the few research on physiological correlates of intonation in English.

Maeda, Shinji. 1982. A digital simulation method of the vocal-tract system. *Speech communication* 1(3-4). 199–229.

Maeda, Shinji. 1990. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech production and speech modelling*, 131–149. Springer.

Marchal, Alain. 2011. *Précis de physiologie de la production de la parole*. Marseille: Solal, "Collection Voix, Parole, Langage".

Martin, James G., Carol B. Mills, Richard H. Meltzer & Joyce H. Shields. 1980. Anticipatory coarticulation and reaction time to phoneme targets in spontaneous speech. *Phonetica* 37(3). 159–168.

Martinet, André. 1945. *La prononciation du français contemporain: Témoignages recueillis en 1941 dans un camp d'officiers prisonniers*. Vol. 3. Librairie Droz.

Martinet, André. 1949. La double articulation du langage. *TCLing-Copenhague* 5. 30–37.

Martinet, André. 1955. *Economie des changements phonétiques*. Berne.

Martinet, André. 1956. *La description phonologique, avec application au parler franco-provençal d'hauteville (savoie)*. Librairie Droz.

Martinet, André. 2020. Economie des changements phonétiques.

Matte, Edouard Joseph. 1982. *Histoire des modes phonétiques du français*. Vol. 162. Librairie Droz.

Maye, Jessica, Daniel J. Weiss & Richard N. Aslin. 2008. Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental science* 11(1). 122–134.

Mcclean, Michael D. & W. R. Tiffany. 1973. The acoustic parameters of stress in relation to syllable position, speech loudness and rate. *Language and Speech* 16(3). 283–290.

McGurk, Harry & John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264(5588). 746–748.

Morton, Eugene S. 2006. Sound symbolism and its role in non-human vertebrate. *Sound Symbolism*. 348.

Näätänen, Risto, Anne Lehtokoski, Mietta Lennes, Marie Cheour, Minna Huotilainen, Antti Iivonen, Martti Vainio, Paavo Alku, Risto J Ilmoniemi, Aavo Luuk, et al. 1997. Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature* 385(6615). 432–434.

Narayanan, Shrikanth S. & Abeer Alwan. 1996. Imaging applications in speech production research. In *Medical imaging 1996: Physiology and function from multidimensional images*, vol. 2709, 120–131.

Nooteboom, Sibout Govert. 1972. *Production and perception of vowel duration: A study of durational properties of vowels in Dutch*. Utrecht University. (Doctoral dissertation).

Ohala, John J. 1983. The origin of sound patterns in vocal tract constraints. In *The production of speech*, 189–216. Springer.

Ohala, John J. 1994. The frequency code underlies the sound-symbolic use of voice pitch. In Leanne Hinton & Johanna Nichols andJohn J. Ohala (eds.), *Sound symbolism*, vol. 2, 325–347.

Ohala, John J. 2011. Sound change is drawn from a pool of synchronic variation. In *Language change*, 173–198. De Gruyter Mouton.

Ohala, John J. 2017. Phonetics and historical phonology. *The handbook of historical linguistics*. 667–686.

Ohala, John J. & Jeri J. Jaeger. 1986. *Experimental phonology*. Academic Press Orlando.

Ohala, John J., Carrie S. Masek, Roberta A. Hendrick & Mary Frances Miller. 1981. The listener as a source of sound change. *parasession on language and behavior (Chicago Linguistics Society, Chicago 1981)*.

Perkell, Joseph S. 1970. Physiology of speech production: Results and implications of a quantitative and cineradiographic study (kim)(book review). *General Linguistics* 10(3). 182.

Perkell, Joseph S., Marc H. Cohen, Mario A. Svirsky, Melanie L. Matthies, Iñaki Garabieta & Michel T. T. Jackson. 1992. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America* 92(6). 3078–3096.

Peterson, Gordon E. & Harold L. Barney. 1952. Control methods used in a study of the vowels. *The Journal of the acoustical society of America* 24(2). 175–184.

Pfleumer, Fritz. 1928. Lautschriftträger, DE000000500900A (auch: Deutsches Reichspatent [DRP] Nr. 500 900). *Reichspatentamt* 5. 1930.

Pickett, James. 1998. *The acoustics of speech communication: Fundamentals, speech perception theory, and technology*. London: Pearson.

Pierrehumbert, Janet. 1979. The perception of fundamental frequency declination. *JASA* 66(2). 363–69.

Pierrehumbert, Janet B. 2003. Phonetic diversity, statistical learning, and acquisition of phonology. *Language and speech* 46(2-3). 115–154.

Pisoni, David B. & Robert E. Remez. 2008. *The handbook of speech perception*. John Wiley & Sons.

Pope, Mildred Katharine. 1934. *From Latin to Modern French with especial consideration of Anglo-Norman: Phonology and morphology*. Manchester University Press.

Price, Cathy Jo. 2000. The anatomy of language: Contributions from functioal neuroimaging. *Journal of Anatomy* 197(3). 335–359.

Prince, Alan & Paul Smolensky. 1993. *Optimality Theory: Constraint interaction in generative grammar*. JManuscript, Rutgers University & University of Colorado, Boulder.

Redford, M. A. (ed.). 2015. *The handbook of speech production*. Oxford/New Jersey: Wiley-Blackwell.

Rogers, Henry. 2014. *The sounds of language: An introduction to phonetics*. 2nd edn. London/New York: Routledge.

Rossi, Mario. 1985. L'intonation et l'organisation de l'énoncé. *Phonetica* 42(2-3). 135–153.

Sapir, Edward. 1925. Sound patterns in language. *Language* 1(2). 37–51.

Schoentgen, Jean. 2006. Vocal cues of disordered voices: An overview. *Acta Acustica united with Acustica* 92(5). 667–680.

Schwartz, Jean-Luc, Louis-Jean Boë, Nathalie Vallée & Christian Abry. 1997. The dispersion-focalization theory of vowel systems. *Journal of phonetics* 25(3). 255–286.

Schweyer, Donald H. 1987. *Delattre's modes: Articulatory settings in applied phonetics for english-speaking learners of french*. University of Alberta.

Selkirk, Elisabeth. 1984. Phonology and syntax: The relation between sound and structure. *Current Studies in Linguistics* I.

Shriberg, Lawrence D., Raymond D. Kent, Tara. McAllister & Jonathan L. Preston. 2018. *Clinical phonetics*. 5th edn. London: Pearson.

Shriberg, Lawrence D., Raymond D. Kent & Benjamin Munson. 2003. *Clinical phonetics*. Boston: Allyn & Bacon.

Sievers, Eduard. 1893. Grundzüge der phonetik.

Silverman, Kim E. A., Mary E. Beckman, John F. Pitrelli, Mari Ostendorf, Colin W. Wightman, Patti Price, Janet B. Pierrehumbert & Julia Hirschberg. 1992. ToBI: A standard for labeling English prosody. In *Icslp*, vol. 2, 867–870.

Sluijter, Agaath M. C. & Vincent J. Van Heuven. 1996. Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical society of America* 100(4). 2471–2485.

Solé, Maria-Josep. 2018. Articulatory adjustments in initial voiced stops in spanish, french and english. *Journal of Phonetics* 66. 217–241.

Stein, Barry E., Terrence R. Stanford, Ramnarayan Ramachandran, Thomas J. Perrault & Benjamin A. Rowland. 2009. Challenges in quantifying multisensory integration: Alternative criteria, models, and inverse effectiveness. *Experimental Brain Research* 198(2-3). 113.

Stemmer, B. & H. Whitaker (eds.). 1998. *Handbook of neurolinguistics*. Cambridge/Massachusetts: Academic Press.

Stemmer, Brigitte & Sieglinde Lacher. 1998. Neurolinguistic and related journal and book resources: A listing. In *Handbook of neurolinguistics*, 641–653. Elsevier.

Stevens, Kenneth N. 1989. On the quantal nature of speech. *Journal of phonetics* 17(1-2). 3–45.

Stevens, Kenneth N. 2000. *Acoustic phonetics*. MIT press.

Stevens, Kenneth N. & Sheila E. Blumstein. 1978. Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America* 64(5). 1358–1368.

Stevens, Kenneth N. & Sheila E. Blumstein. 1981. The search for invariant acoustic correlates of phonetic features. *Perspectives on the study of speech*. 1–38.

Stevens, Kenneth N. & Arthur S. House. 1955. Development of a quantitative description of vowel articulation. *The Journal of the Acoustical Society of America* 27(3). 484–493.

Stone, Maureen. 1997. Laboratory techniques for investigating speech articulation. *The handbook of phonetic sciences* 1. 1–32.

Straka, Georges. 1965. *Album phonétique*. Presses de l'Université Laval.

Studdert-Kennedy, Michael & Kerstin Hadding. 1973. Auditory and linguistic processes in the perception of intonation contours. *Language and Speech* 16(4). 293–313.

Sun-Ah, Jun (ed.). 2005. *Prosody typology*. Vol. 1. Oxford, Linguistics.

Sun-Ah, Jun (ed.). 2014. *Prosody typology*. Vol. 2. Oxford, Linguistics.

Sundberg, Johan, Lennart Nord & Rolf Carlson. 1990. *Music, language, speech and brain: Proceedings of an International Symposium at the Wenner-Gren Center, Stockholm, 5–8 September 1990*. Macmillan International Higher Education.

Syrdal, Ann K., Raymond W. Bennett & Steven L. Greenspan. 1994. *Applied speech technology*. CRC press.

Thorsen, Nina G. 1980. A study of the perception of sentence intonation: Evidence from Danish. *The Journal of the Acoustical Society of America* 67(3). 1014–1030.

Titze, Ingo R. & Daniel W. Martin. 1998. *Principles of voice production*.

Tomás, Tomás Navarro. 1974. *Manual de entonación española*. Vol. 175. Guadarrama.

Torre III, Peter & Jessica A Barlow. 2009. Age-related changes in acoustic characteristics of adult speech. *Journal of communication disorders* 42(5). 324–333.

Troubetzkoy, Nikolaï Sergueïevitch. 1949. Principes de phonologie, trad. *Cantineau J., Klincksieck,(réimpression. 1964), Paris*.

Vaissière, Jacqueline. 1994. Phonological use of the larynx: A tutorial. *Larynx* 97. 115–126.

Vaissiére, Jacqueline. 1971. *Contribution à la synthèse par règles du français*.

Vaissiére, Jacqueline. 2007. Area functions and articulatory modeling as a tool for investigating the articulatory, acoustic and perceptual properties of sounds across languages. *Experimental approaches to phonology*. 54–71.

Vaissière, Jacqueline. 1983. Language-independent prosodic features. In *Prosody: Models and measurements*, 53–66. Springer.

Vaissière, Jacqueline. 1991. Rhythm, accentuation and final lengthening in french. In *Music, language, speech and brain*, 108–120. Springer.

Vaissière, Jacqueline. 1995. Phonetic explanations for cross-linguistic prosodic similarities. *Phonetica* 52(3). 123–130.

Vaissière, Jacqueline. 1996. From Latin to Modern French: On diachronic changes and synchronic variations. *AIPUK, Arbetisberitche, Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel* (31). 61–74.

Vaissière, Jacqueline. 2001. Changements de sons et changements prosodiques: Du latin au français. *Revue parole* (17). 53–88.

Vaissière, Jacqueline. 2007. Area functions and articulatory modeling as a tool for investigating the articulatory, acoustic and perceptual properties of sounds across languages. *Experimental approaches to phonology*. 54–71.

Vaissière, Jacqueline. 2008. Perception of intonation. *The handbook of speech perception*. 236–263.

Vaissière, Jacqueline. 2011. On the acoustic and perceptual characterization of reference vowels in a cross-language perspective. In *The 17th international congress of phonetic sciences (icphs xvii)*, 52–59.

Vallée, Nathalie. 1994. *Systèmes vocaliques: De la typologie aux prédictions*. Grenoble 3. (Doctoral dissertation).

Visconti, Jacqueline (ed.). 2018. *Handbook of communication in the legal sphere*. Germany: De Gruyter.

Von Helmholtz, Hermann. 1867. *Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln*. Vol. 9. Voss.

Warren, D. W. & D. J. Hall. 1973. Glottal activity and intraoral pressure during stop consonant productions. *Folia Phoniatrica et Logopaedica* 25(1-2). 121–129.

Warren, Richard M. & Gary L. Sherman. 1974. Phonemic restorations based on subsequent context. *Perception & Psychophysics* 16(1). 150–156.

Werker, Janet F. & Suzanne Curtin. 2005. PRIMIR: A developmental framework of infant speech processing. *Language learning and development* 1(2). 197–234.

Westbury, John R. 1983. Enlargement of the supraglottal cavity and its relation to stop consonant voicing. *The Journal of the Acoustical Society of America* 73(4). 1322–1336.

Whalen, Douglas H. 1984. Subcategorical phonetic mismatches slow phonetic judgments. *Perception & psychophysics* 35(1). 49–64.

Windsor, Fay, M. Louise Kelly & Nigel Hewlett. 2012. *Investigations in clinical phonetics and linguistics*. Psychology Press.

Yeou, Mohamed & Shinji Maeda. 1995. Pharyngeal and uvular consonants are approximants: An acoustic modeling study. In *Proceedings of the 13th international congress of phonetic sciences*, 586–589.

Yu, Dong & Li Deng. 2016. *Automatic speech recognition*. Springer.

Zhou, Xinhui, Carol Y Espy-Wilson, Suzanne Boyce, Mark Tiede, Christy Holland & Ann Choe. 2008. A magnetic resonance imaging-based articulatory and acoustic study of “retroflex” and “bunched” American English/r. *The Journal of the Acoustical Society of America* 123(6). 4466–4481.

Ziegler, Wolfram. 2008. 39. assessment methods in neurophonetics: Speech production. In *Linguistic disorders and pathologies*, 432–443. De Gruyter Mouton.

Zue, Victor W. 1985. The use of speech knowledge in automatic speech recognition. *Proceedings of the IEEE* 73(11). 1602–1615.

Zue, Victor W. & Martha Laferriere. 1979. Acoustic study of medial/t, d/in american english. *The Journal of the Acoustical Society of America* 66(4). 1039–1050.

Phonetic sciences

Set blurb on back with \BackBody{my blurb}