INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

ma per seguir virtute e canoscenza

SISSA

Cross-linguistic exploration of phonemic representations

Candidate: Zeynep Gökçen Kaya

Advisor: Alessandro Treves

Thesis submitted for the degree of Doctor of Philosophy in Neuroscience

Trieste, 2018

Contents

List of Figures

List of Tables

Abstract

All languages around the world have their own vast sound inventories. Understanding each other through verbal communication requires, first of all, understanding each other's phonemes. This often overlooked constraint is non-trivial already among native speakers of the same language, given the variability with which we all articulate our phonemes. It becomes even more challenging when interacting with non-native speakers, who have developed neural representations of different sets of phonemes. How can the brain make sense of such diversity?

It is remarkable that the sounds produced by the vocal tract, that have evolved to serve as symbols in natural languages, fall almost neatly into two classes with such different characteristics, consonants and vowels. Consonants are complex in nature: beyond acoustically-defined formant (resonant) frequencies, additional physical parameters such as formant transitions, the delay period in those transitions, energy bursts, the vibrations of the vocal cords occurring before and during the consonant burst, and the length of those vibrations are needed to identify them. Surprisingly, consonants are very quickly categorized through a quite mysterious form of invariant feature extraction. In contrast to consonants, vowels can be represented in a simple and transparent manner and that is because, amazingly, only two analog dimensions within a continuous space are essentially enough to characterize a vowel. The first dimension corresponds to the degree to which the vocal tract is open when producing the vowel and the second dimension is the location of the main occlusion. Surprisingly, these anatomically-defined production modes match very precisely the first two acoustically-defined formant frequencies, namely F1 and F2. While for some languages some additional features are necessary to specify a vowel, such as its length or roundedness, whose nature may be more discrete, for many others F1 and F2 are all there is to it.

In this thesis, we use both behavioral (phoneme confusion frequencies) and neural measures (the spatio- temporal distribution of phoneme-evoked neural activation) to study the cross-linguistic organization of phoneme perception. In Chapter 2, we study the perception of consonants by replicating and extending a classical study on sub-phonemic features underlying perceptual differences between phonemes. Comparing the responses of native listeners to that of Italian, Turkish, Hebrew, and (Argentinian) Spanish listeners to a range of American English consonants, we look at the spe-

cific patterns of errors that speakers of different languages make by using the metric content index, which was previously used in entirely different contexts, with either discrete, e.g. in face space, or continuous representations, e.g. of the spatial environment. Beyond the analysis of percent correct score, and transmitted information, we frame the problem in terms of ‘place attractors’, in analogy to those which have been well studied in spatial memory. Through our experimental paradigm, we try to access distinct attractors in different languages. In the same chapter, we provide auditory evoked potentials of some consonant-vowel syllables, which hint at transparent processing of the vowels regulated by the first two formants that characterize them, and accordingly we then turn to investigating the vowel trajectories in the vowel manifold.

We start our exploration of the vowel space in Chapter 3 by addressing a perceptually important third dimension for native Turkish speakers – that is rounding. Can native Turkish speakers navigate better vowel trajectories in which the second formant changes over a short time, to reflect rounding, compared to native Italian speakers, who are not required to make such fine discriminations on this dimension? We found no mother tongue effects. We have found, however, that rounding in vowels could be represented with similar efficiency by fine differences in a F2 peak frequency which is constant in time, or inverting the temporal dynamics of a changing F2, which then makes vowels not mere points in the space, but rather continuous trajectories.

We walk through phoneme trajectories at every tens of milliseconds, it comes to us as naturally as walking in a room, if not more. Similar to spatial trajectories, we create equidistant continuous vowel trajectories in Chapter 4 on a *vowel wheel* positioned in the central region of the two-dimensional vowel space where in some languages like Italian there are no standard vowel categories, and in some other, like English, there are. Is the central region in languages like Italian to be regarded as a flat empty space with no attractors? Is there any reminiscence of their own phoneme memories? We ask whether this central region is flat, or can at least be flattened through extensive training. If so, would then we find a neural substrate that modulates the perception in the 2D vowel plane, similar to grid cell representation that is involved in the spatial navigation of empty 2D arenas? Our results are not suggestive of a grid-like representation, but rather points at the modulation of the neural signal by the position of Italian vowels around the outer contour of the wheel.

Therefore in Chapter 5, we ask how our representation of the vowel space, not only in the central region but rather in the entirely of its linguistically relevant portion, is deformed by the presence of the standard categories of our vowel repertoire. We use ‘belts’, that are short stretches along which formant frequencies are varied quasi-continuously, to determine the local metric that best describes, for each language, the vowel manifold as a non-flat space constructed in our brain. As opposed to the ‘consonant planes’, that we constructed in Chapter 2, which appear to have a similar structure to a great extent, we find that the vowel plane is subjective and that it is language dependent. In light of language-specific transformations of the vowel plane, we wonder whether native bilinguals hold simultaneously multiple maps available and use one or the other to interpret linguistic sources depending on context. Or alternatively, we ask, do they construct and use a fusion of the two original maps, that allows them to efficiently discriminate vowel contrast that have to be discriminated in either language? The neural mechanisms underlying the physical map switch, known as remapping, have been well studied in rodent hippocampus; is the vowel map alternation governed by similar principles? We compare and show that the perceptual vowel maps of native Norwegian speakers, who are not bilingual but fluent in English, are unique, probably sculpted by their long-term memory codes, and we leave the curious case of bilinguals for future studies.

Overall we attempt to investigate phoneme perception in a different framework compared to how it has been studied in the literature, which has been in the interest of a large community for many years, but largely disconnected from the study of cortical computation. Our aim is to demonstrate that insights about persisting questions in the field may be reached from another well explored part of cognition.

Acknowledgements

I would like to acknowledge the contributions of various people in different parts of this thesis; Yair Lakretz for the data collection of Hebrew participants for the experiments in Chapter 2 and for his help with creating the auditory stimuli for the experiments in Chapter 5 together with Joe Collins, who also collected the data of Norwegian participants, Simona Perrona, who collected the data of Scottish participants in Chapter 5, Yamil Vidal Dos Santos, who provided the clustering analysis pipeline in Chapter 4, Mohammad Reza Soltanipour, for his major help and work with Fourier decomposition and PCA analysis of EEG data in Chapter 4, Simona Sausa for her help in stimuli creation and participant recruitment in Chapter 3, and Alessio Isaja for his help with the technical problems in the EEG lab.

I would like to express my gratitude to Professor İlknur Çavuşoğlu in Faculty of Medicine of Uludağ University for hosting me to conduct all of the experiments with Turkish speakers, and to Professor Alex Gomez-Marin in Instituto de Neurociencias (Alicante) for hosting me to collect the data of Spanish participants in Chapter 5.

I would first like to thank my supervisor Alessandro Treves for his persistence in turning me into a phonologically aware person, and for his continuous support.

I would like to thank my fellow lab-mates for the time we spent together in LIMBO, but especially to Massimiliano (Max), for his kind help and availability whenever I asked for. I would like to thank my friends with whom I have enjoyed Trieste together; Vizhe, Georgette, Yamil, and Romain for our shared understanding in different time-scales, to Reza again, for his presence during the last summer I spent in Trieste, and to Temmuz for making me feel at home from long distance, and my grandparents and parents for their support.

I thank to all genuinely kind people I met over the years including the ones who participated in the experiments, the ones who donated our research their phonemes, and the ones who helped to solve daily problems like Riccardo Iancer.

Finally, I cannot express my gratitude to Giangiacomo enough for making everything far more enjoyable in an unexpected way with his unprecedented patience, generosity, love, support and

intellectual curiosity intertwined with bad humor taste.

1

General Introduction

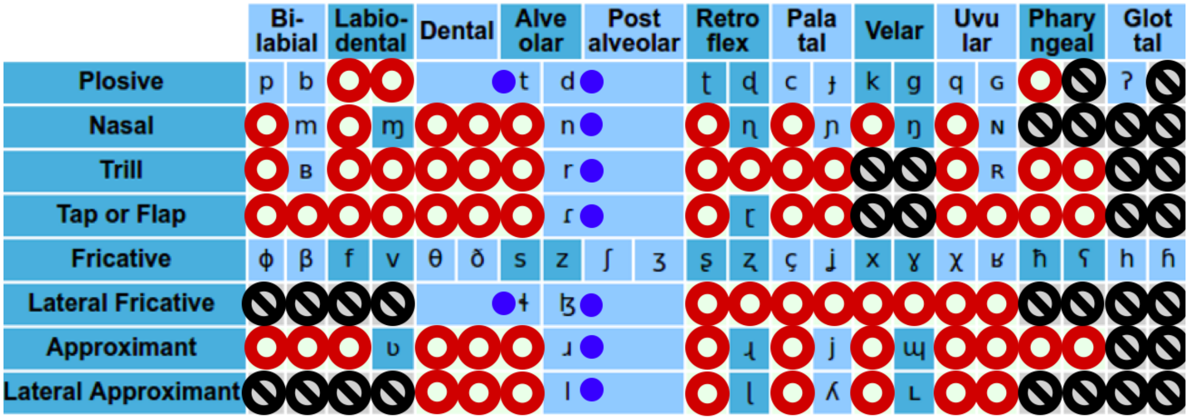
As we speak, we produce streams of highly variable and complex sound patterns. How can the human brain perceive and process these fleeting speech signals and map them to some linguistic representation? At the lowest level, this requires mapping the acoustic signal into sound units called *phonemes*, a process which makes us differentiate first syllables and then words from each other. For example, one−syllable words fork (/førk/) and cork (/cørk/) have four phonemes and we recognize them as distinct words because their initial phonemes are not the same, although the rest is.

Phonemes are classified either as vowels or consonants according to their spectral properties. Frequency analyses of a vowel sound always show a clear harmonic spectrum determined by the vibration of the vocal folds with several components, each with its own amplitude and bandwidth. The first one is the *pitch* (fundamental frequency i.e. F0), which varies from individual to individual and is influenced by gender. Female speakers usually have higher pitch than male speakers. The acoustic correlates of most consonants do not reveal harmonic frequency spectra. The characteristic features of consonant sound patterns are a mixture of operations like energy bursts that resemble white noise, periods of silence, voicing.

While the speech signal is continuous, the phonetic description is alway discrete. As in the /førk/ and /cørk/ examples, phonemes are transcribed by an international system of alphabetic notation (IPA) [IPA, 2018]. This notation is designed to comprise all qualities included in the oral language. Next we will recapitulate how the two sets of sounds are represented following the traditional notation and the factors governing their perception.

1.1. Consonants

The conventional representation of consonants is given by a simple grid that emerges from the combination of two sub-phonemic features: *place of articulation* and *manner of articulation*, as shown in Fig. 1.1. Place of articulation refers to the physical point in the vocal tract where the obstruction occurs, as shown by the columns of the grid. Places along the vocal tract are laid out starting from the most rostral part, i.e. the lips, to the most caudal of it, i.e. the glottis. Notice, that there is

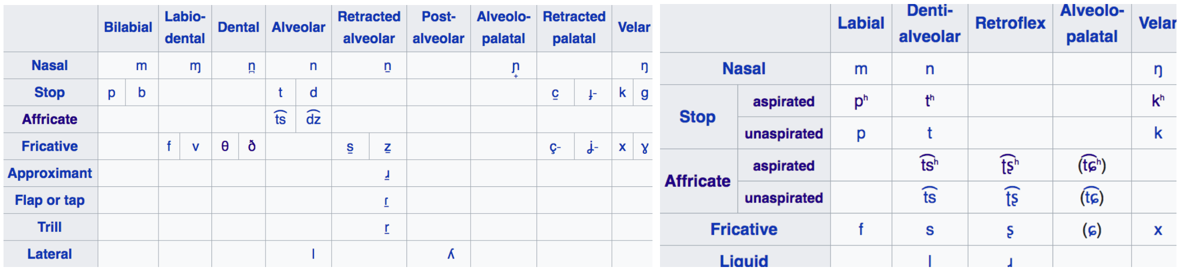


	Bi- labial		Labio- dental		Dental		Alve olar		Post alveolar		Retro flex		Pala tal		Velar		Uvu lar		Phary ngeal		Glot tal	
Plosive	p	b	○	○			●t	d●			ʈ	ɖ	c	ɟ	k	g	q	ɢ	○	⊘	ʔ	⊘
Nasal	○	m	○	ɱ	○	○	○	n●			○	ɳ	○	ɲ	○	ŋ	○	ɴ	⊘	⊘	⊘	⊘
Trill	○	ʙ	○	○	○	○	○	r●			○	○	○	○	⊘	⊘	○	ʀ	○	○	⊘	⊘
Tap or Flap	○	○	○	○	○	○	○	ɾ●			○	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	○	○	⊘	⊘	⊘	⊘

Fig. 1.1: Pulmonic consonants imposed on a grid. The traditional representation of consonants is given by a 2D grid table although the space of all possible consonant vocalizations does not correspond to continuous manifold. Red circles denote combinations that humans are physically capable of producing (but are not normally produced), black banned signs are the ones which are impossible. The columns and rows of the table correspond to place and manner of articulation respectively. Blue circles show the consonants that do not have a fixed place of articulation. Most place columns are split in two, with voiceless phonemes to the left and voiced ones to the right. Note that the manner dimension has an arbitrary arrangement unlike the place dimension, which follows the position along the vocal tract. Not all manners are embedded in the table, and for each language a different arrangement representation can be proposed, as shown in Fig. 1.2.

no definite place of articulation attributed to every sound, as noted by the blue circles in Fig. 1.1. Some sounds can change their place of articulation depending on their position in a word and from one language to another.

The way articulatory organs interact with each other during consonant production is described as the manner of articulation. Although different manners are laid out as shown in the rows of Fig. 1.1, like nasals coming before trills and after plosives, this arrangement is artificial and does not correspond to a physical reality, unlike that for place articulation.



	Bilabial		Labio-dental		Dental		Alveolar		Retracted alveolar		Post-alveolar		Alveolo-palatal		Retracted palatal		Velar	
Nasal		m		ɱ		n̪		n		n̠				ɲ̟				ŋ
Stop	p	b					t	d							c̠	ɟ˗	k	g
Affricate							t͡s	d͡z										
Fricative			f	v	θ	ð			s̠	z̠					ç˗	ʝ˗	x	ɣ
Approximant										ɹ̠								
Flap or tap										ɾ̠								
Trill										r̠								
Lateral								l				ʎ						

		Labial	Denti-alveolar	Retroflex	Alveolo-palatal	Velar
Nasal		m	n			ŋ
Stop	aspirated	pʰ	tʰ			kʰ
	unaspirated	p	t			k
Affricate	aspirated		t͡sʰ	t͡ʂʰ	(t͡ɕʰ)	
	unaspirated		t͡s	t͡ʂ	(t͡ɕ)	
Fricative		f	s	ʂ	(ɕ)	x
Liquid			l	ɹ		

Fig. 1.2: 2D representations of consonants in 2 different languages. Different languages have different consonants. Here on the left panel, the arrangement of Greek consonants, and on the right that of Chinese consonants are given. Voicing (but not aspiration) is perceptually salient for native Greek speakers whereas aspiration (but not voicing) is contrastive for native Chinese speakers. In Greek, nasal sounds produced at almost all adjacent points are realized as a distinct phoneme. In Chinese, 7 different sounds are produced with a flat tongue against the alveolar ridge and upper teeth (i.e. denti-alveolar). Further differences between the two can be observed.

Notice that, for plosives (also called stops) and fricatives, sounds produced at a given place of articulation come in pairs by the voicing feature which arises from the laryngeal configuration. When the larynx is closed, the vocal folds inside the glottis vibrate and the sound is voiced, such as for the labio-dental fricative /v/. When the larynx is open, the sound is only produced by the burst or the turbulence at the place of constriction, and therefore it is voiceless like for the labio-dental fricative /f/. In fact, the laryngeal configuration can have a further complexity during the production of plosives, that results in a different time difference between the release of a plosive and the onset of the vibration of the vocal folds. That duration is called *voiced-onset timing* (VOT) which can be an important perceptual cue for plosive recognition [Eimas and Corbit, 1973, Laufer, 1998]. On the other hand, approximant and nasal sounds are always voiced. Although there are other manner perceptual cues, they are usually not embedded in the grid representation. Chinese, for example, has stops and affricates distinguished by *aspiration*, that is the strong burst of breath, i.e. glottal turbulence, following the release of the consonant (see the right panel of Fig. 1.2). In Korean, instead, the duration of the oral pressure buildup accompanying the release is discriminative, i.e. a distinction between *tense* and *plain* consonants exists.

Perhaps more than the arrangement of the manner dimension, what puts the representation of consonants further away from the idea of a continuous space is the holes that catch the eye in the grid. The red circles denote the place and manner combinations that articulatory organs are capable of producing, but for which no symbol is assigned. The banned signs, on the other hand, show the combinations that are physically not possible. Therefore, considering all the points we discussed about the Fig. 1.1, the space of consonants is more complex than the one imposed on a simple grid, even though such representation is extremely useful and has persisted over centuries. The earliest example of a similar arrangement of Japanese phonemes, which derives from a much earlier Sanskrit arrangement, dates back dates back to the 13th century, and it has not been further revised since the 17th century [Buck, 1970]. As we will see next, the space of vowels is a different story.

1.2. Vowels

The place and manner system that is used to represent the consonants is adapted also for the representation of vowels, as shown on the left panel of Fig. 1.3. However the crucial difference between the two representations is that these two dimensions, place and manner, can be directly mapped from the formant (vocal tract resonance) frequencies of the acoustic signal of vowels. Manner, i.e. the openness of the vocal tract, and place, i.e. the location of the main occlusion, are expressed by the first and second formant frequencies F1 and F2 respectively, as it is seen in the right panel of Fig. 1.3. Notice the correspondence between the positions of the vowels /i/, /e/, /ɛ/, /æ/ in the 2D plane in the left panel, and the increase in the F1 and the decrease in the F2 values in their respective spectrograms going from /i/ to /æ/. The more open and the more front a vowel is, the bigger are the first formant and the second formant values. Some languages can use extra dimensions, like

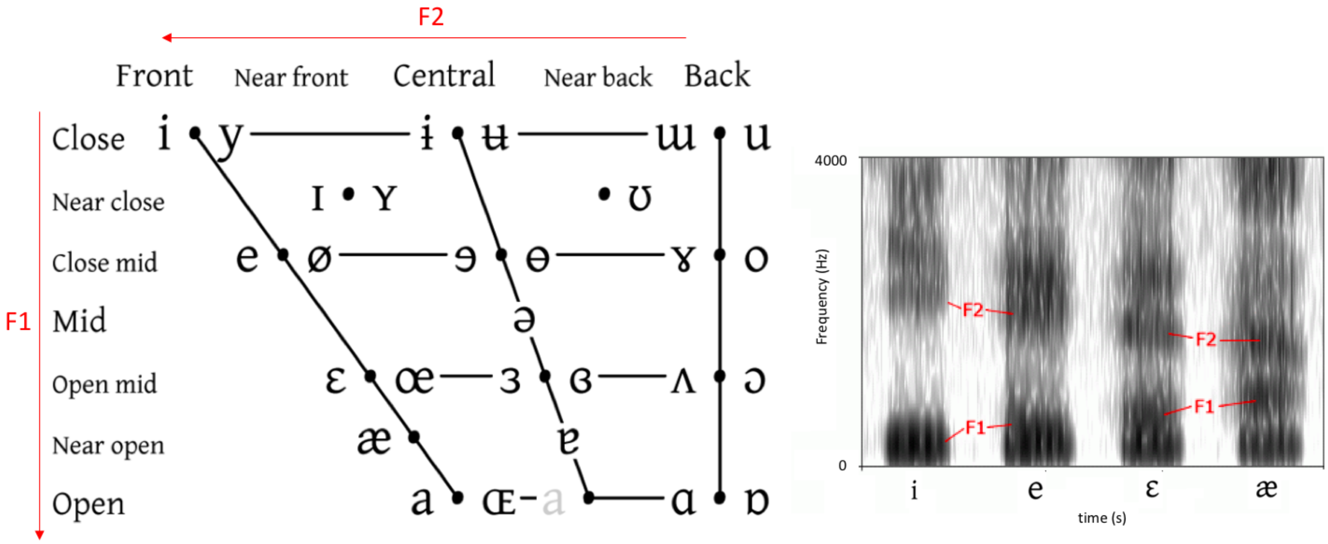


Fig. 1.3: The manifold nature of vowels. Vowels can be described by the interaction of two dimensions: openness of the vocal tract (manner) and the position of the main occlusion in the vocal tract (place). Amazingly, as opposed to consonants, these two features directly translate into the acoustic signal and reflect to the first and second formant (resonant) frequencies respectively, as shown with the spectrograms of four vowels (/i/ /e/ /ɛ/ /æ/) on the right panel. The first formant (F1) increases as the vowel changes from close to open, depending on the degree of openness of the vocal tract. The second formant (F2) increases as the articulation occurs at a more frontal portion of the oral cavity. Given this transparent mapping, vowels can be described by fixed points or short trajectories on a two-dimensional manifold. Adapted from [Russell, 2009].

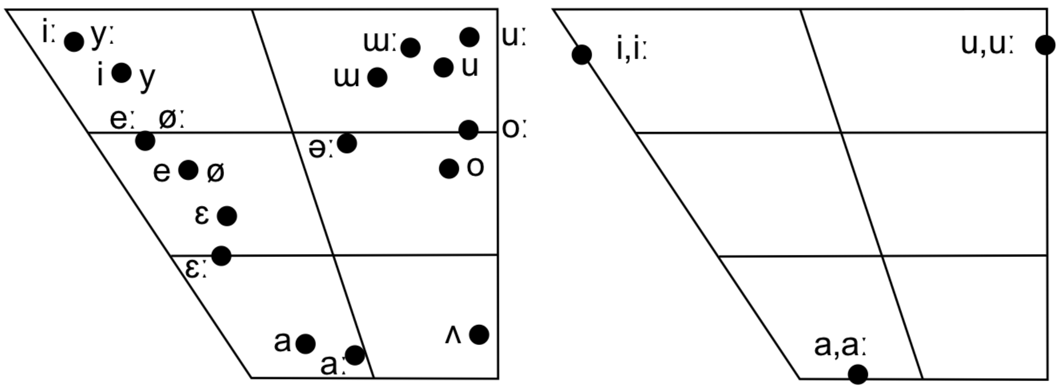


Fig. 1.4: The vowel spaces of 2 different languages. Languages differ in the number and the position of their vowels. The Korean vowel inventory, on the left, is rich; small steps taken on the plane can end at different vowel categories. In (classical) Arabic, there are only three vowels but with long-short contrast for each one of them. Presumably, such differences change the local geometry of each plane.

rounding. In the left panel of Fig. 1.3 the acoustic correlates of rounding is captured within the F2 dimension, where among a pair of vowels the one that is relatively more back is the rounded sound. For a given language, another contrastive feature can be *duration*, which is simply the length of the signal. For instance, Arabic and Korean are among the languages for which duration is a phonemic perceptual cue, as shown in Fig. 1.4, where the vowels with the symbol ː denote the long ones. From the same figure, we can infer that phoneme inventories of languages differ in their vowels as they do in their consonants. The same vowel can have different positions on the language-specific plane (see the position of /i/, for example). As shown in Fig. 1.5, even within a given language, the position of a vowel can vary greatly, depending for example on individual physical differences like gender, head size, vocal tract length, or even emotional mood, all of which we experience during natural speech.

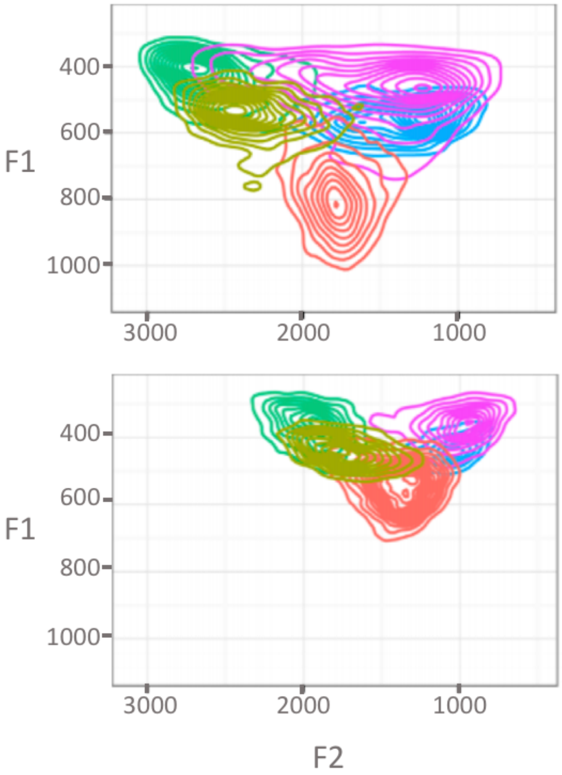


Fig. 1.5: Vowel variability during natural speech. The dispersion of five vowels /i/ (dark green) /e/ (light green) /a/ (red) /o/ (blue) /u/ (pink) during spontaneous speech by three female speakers (on the top panel) and four male speakers (on the bottom panel). Formant values are drawn as two-dimensional contours. Original image from [DiCanio et al., 2015].

Vowels are also called monophthongs. If the tongue moves during vowel production from one position towards another, it results in the production of one syllable with two vowels, which are together called diphthongs, which literally means “two sounds” in Greek. However, what is a diphthong and whether its realization is one single phoneme or two adjacent phonemes lies in the cognition of the listeners [Balas, 2009]. For example /ai/ can be recognized as a diphthong in a given language, but the reverse i.e. /ia/ does not have to be necessarily a diphthong. As seen from the spectrograms in Fig. 1.6, a diphthong has at least one steady and one transitioning part [Fox, 1983]. On the 2D plane in Fig. 1.3, they can be represented as trajectories. The slope and the duration of the transitioning part can be discriminative perceptual cues [Gay, 1970].

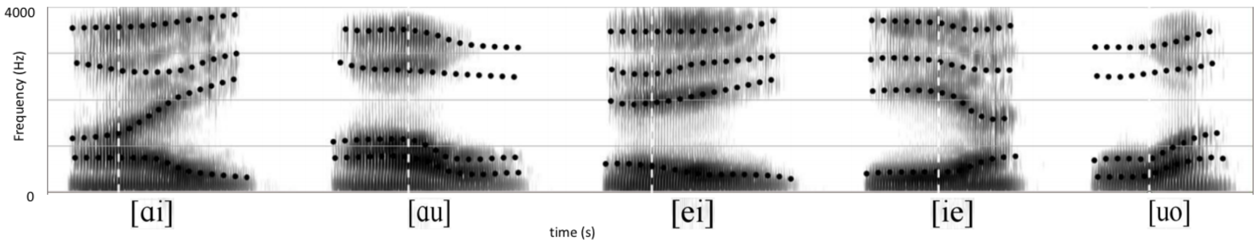


Fig. 1.6: Lithuanian diphthongs. Diphthongs differ from monophthongs as their structure reveals at least one steady and one transitioning part. Whether each member of the pair is recognized as one phoneme or not depends on the language. Notice that the reverse of /ei/ is also a diphthong in Lithuanian, but the reverse of /uo/, for example, is not. Original image taken from [Sproge, 2003].

The structure of these systems of phonemes, consonants and vowels, that we have briefly reviewed here, is expected to be reflected in their representation in the brain, and to influence our linguistic behavior. Exactly how, is the question or rather set of questions that we approach with the studies presented in this Thesis.

1.3. Processing consonants vs. vowels

Spatial patterns of cortical activity during speech production follow dynamic trajectories, as shown in Fig. 1.7. How is the difference in the production of consonants and vowels reflected in their phonological representation? Are they processed as distinct ‘entities’ or are they just two different facets of the smallest unit of language processing?

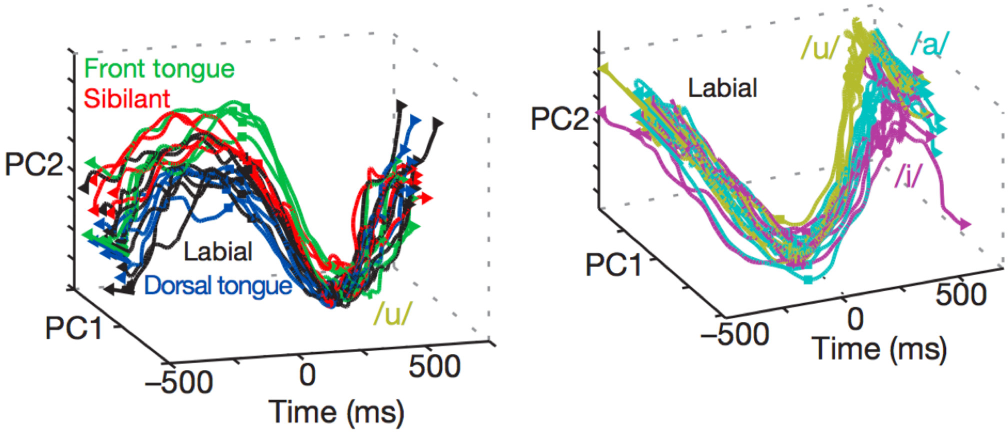


Fig. 1.7: Cortical state-space during speech production. (Left panel) Neural trajectories of consonant-vowel syllables during articulation, as extracted from multi-electrode cortical recordings. Different consonants are transitioning into vowel /u/. For each trajectory, the left triangle indicates 500ms before transitioning into the vowel, the square indicates 25ms before, the circle indicates 250ms after the transition and the right triangle indicates 750ms after. (Right panel) The state-space when labial consonants are transitioning into three different vowels. Original image taken from [Bouchard et al., 2013].

The performance of aphasic patients' at repeating words shows that consonants and vowels can be damaged independently [Caramazza et al., 2000]. Similarly, the modulation of the stimulation of left superior temporal gyrus impairs discrimination of consonants but not vowels [Boatman et al., 1997]. The authors reasoned that one explanation for the consonant-vowel dissociation could be the necessity for cue-integration for consonant perception, but not for vowel perception, for which single acoustic cues, like steady-state formant frequency values are enough.

A further insight is brought forth by testing the recognition memory ('same' or 'different' responses) for pairs of synthetic phonemes, steady-state vowels or consonants, to be discriminated from either 'between' or 'within' phonetic categories, by varying the length of the comparison interval from zero to two seconds [Pisoni, 1973]. For consonants and long vowels (i.e. 300ms long) between category performance was high and independent of delay interval. In contrast, short vowels (i.e. 50ms long) were perceived more in a continuous mode so that also within category performance was relatively high. Moreover performance decreased inversely proportional to delay interval, suggesting that it is based on a short-term memory trace, as opposed to the stable encoding of vowel and consonant categories.

A more recent paper gives us an idea of the short-term neural traces left by the acoustic signals i.e. the phoneme-related potentials (PRPs) [Khalighinejad et al., 2017]. Results showed complex waveforms across multiple electrodes, with the power to discriminate among 30 phonemes uttered in continuous speech by two speakers. The average PRPs cluster reasonably in 4 major categories: vowels, approximants, fricatives and plosives. The use of two speakers, one male and one female, illustrates the degree of variability in the neural response to the same phoneme category. The variability appears quite remarkable, especially when focusing on vowels only, as shown in Fig. 1.8 . Thus, it appears that listeners are able to use the appropriate speaker-dependent map to decode early acoustic responses into appropriate phonological categories, even though the neural population expressing them were the same for the two speakers. This suggests a form of relative encoding, e.g. of position in (F1,F2) space, with reference to a speaker-defined origin or landmark. Note that the authors state that the degree of separation of the PRP to the two speakers is virtually zero if considering all PRPs, while if restricted to vowels it is as high as that among manners of articulations (the 4 clusters mentioned above) of all phonemes. No results are provided regarding speaker separation for consonants only, or for each of the 3 consonant clusters.

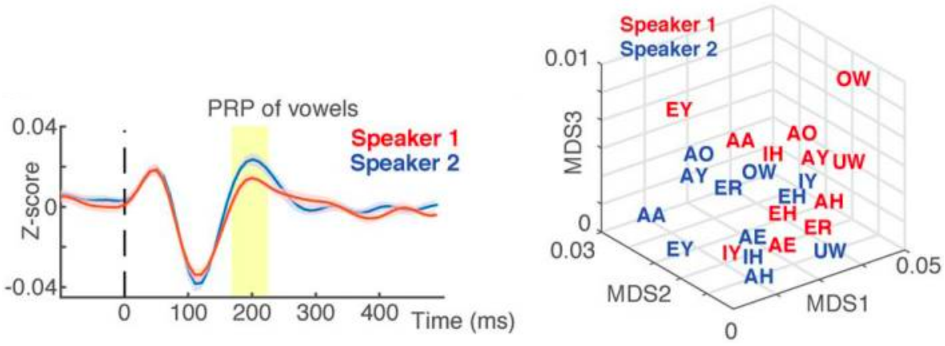


Fig. 1.8: Speaker-dependent vowel reprsentation. (Left panel) Average PRPs to vowels of two speakers (male in blue, female in red) show a difference around 200ms. (Right panel) Multi-dimensional scaling of phoneme-related EEG potentials showing separation of speakers. Original image taken from [Khalighinejad et al., 2017].

As a summary of the evidence on the input representations of phonemes, it appears that both for consonants and for vowels they ‘culminate into’ categorical representations which can last in memory for a second or a little more. Before reaching that stage, consonants are coded by dynamical trajectories, very quickly categorized through an amazingly invariant feature extraction; whereas vowels can persist in a more transparent, static and continuous code, and this may be crucial for extracting from the continuum appropriate categories with different speakers, and most likely with different languages as well, for which the relevant categories differ.

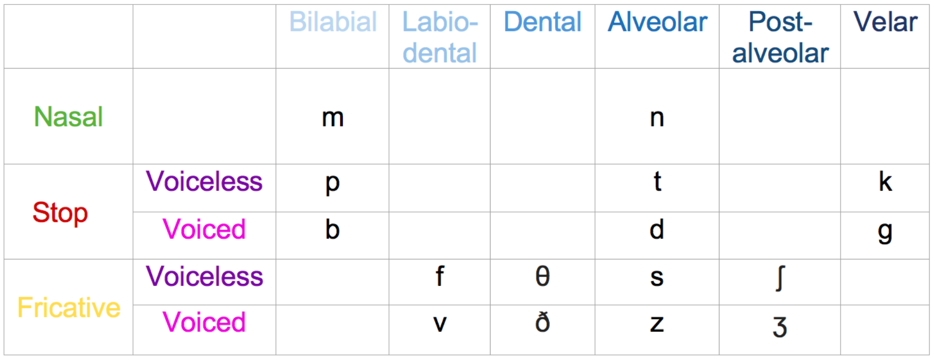
2

Consonants

2.1. The cross–linguistic organization of consonant perception

Often our verbal communications take place in suboptimal conditions in which we have to infer the phonemes of the other; she might say rat (/ræt/) and if there is no contextual information helping us to identify the word as /ræt/, we might hear bat (/bæt/) or cat (/kæt/) as well since their phonetic representation differ from what she said only by the word-initial phoneme. In the seminal study by Miller and Nicely, they used this everyday phenomenon to investigate the structure underlying consonant confusion [Miller and Nicely, 1954]. Native (American) English speakers were presented one of the 16 English consonants (see Fig. 2.1) masked with white noise and they were asked to identify the consonant correctly. The noise naturally gives listeners difficulty in identifying the sound they hear, and makes them confuse sounds with one another. The confusion frequencies of the participants give us a measure of the perceptual similarity of the consonants, so that sounds that are represented closer in mental space are confused more with each other, in a sense forming clusters.

Which perceptual cue characterizes the confusion clusters the most? Do we rely more on the manner or on the place of articulation of the consonants to identify them? Analyzing the transmitted information between the stimulus and response, they showed that, among three manners of articulation, voicing and nasality features were found to be very resistant to noise compared to affrication.



		Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Velar
Nasal		m			n		
Stop	Voiceless	p			t		k
	Voiced	b			d		g
Fricative	Voiceless		f	θ	s	ʃ	
	Voiced		v	ð	z	ʒ	

Fig. 2.1: 16 consonants in English. 16 consonants, which make up the two-thirds of the English consonant inventory, imposed on a grid representation where rows represent manner categories and columns represent places of articulation.

However, information about the place of articulation was the least preserved. This observation was valid across different SNR conditions, even when the signal-to-noise ratio was very low. Multidimensional scaling of functional magnetic resonance imaging (fMRI) data of the same 16 consonants also showed an arrangement of BOLD activity according to the voicing and nasality features [Arsenault and Buchsbaum, 2015].

This experiment was so simple and elegant in its design and analysis that, over more than 70 years, it has been replicated several times. These later studies focused mainly on the effect of the properties of the mask noise (some of which are [Phatak and Allen, 2007,Redford and Diehl, 2004]) or perceptual differences between native and non-native speakers (some of which are [Cutler et al., 2004, Cooke et al., 2010]). Non-native phoneme perception is under the influence of the native phoneme contrasts [Flege, 1995]. Separate studies confirmed the superiority of native listeners at identifying the phonemes of their mother tongue and similar feature sensitivity by native and non-native speakers. Beyond the analyses of percent correct score and mutual information, we extend studying the native vs. non-native differences by looking at the specific patterns of errors for speakers of five different languages.

2.2. Materials and methods

2.2.1. Auditory Stimuli

As in the original study, we have recorded 5 female native American English speakers articulating 16 English consonants, which make up more than half of the English consonant phonemes, spoken in the initial position before the vowel /a/. Four exemplar waveforms and their corresponding spectrograms are shown in Fig. 2.2. All the (16*5= 80) stimuli were first normalized to the same RMS using Praat [Boersma and Weenik, 2018], and then white noise was added so that the final signal-to-noise ratio would be $0dB$, which gives listeners difficulty in identifying the original speech and makes them confuse phonemes with each other.

2.2.2. Participants

We have tested five groups of participants of different mother tongues; Italian, (American) English, Turkish, (Argentinian) Spanish, and Hebrew. In each language group, there were 10 listeners and all the listeners could speak English, but perhaps with varying degrees of their competence. Another group of 15 native Italian listeners was tested on the same stimuli, but with an SNR ratio of $-6dB$.

2.2.3. Paradigm

During the experiment, each listener went through a training phase and a test phase before the experiment phase. During training, they were asked to get themselves acquainted with the 16 phonetic symbols and the sounds they represent, not all of which exist in all the language groups we tested. Once they felt confident enough, they were given seven questions in each of which they listened to a phoneme and chose the corresponding symbol. These questions included two dental (/θ/,/ð/) and two post-alveolar (/ʃ/,/ʒ/) fricative phonemes, denoted by symbols which are less familiar to many of the listeners, and three random ones from the remaining twelve consonants, to make sure the participants learned the associations between the phonetic symbols and the sounds.

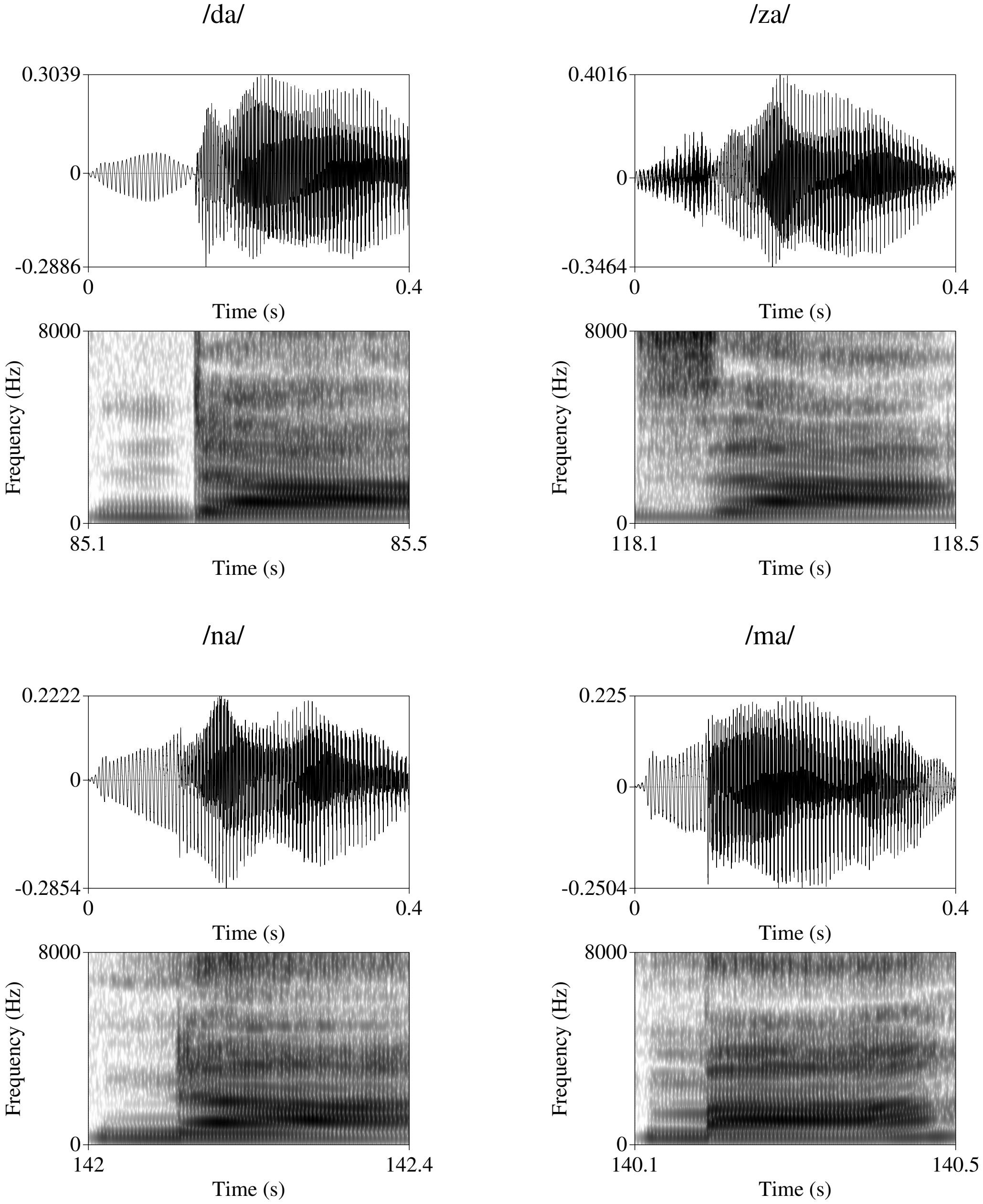


Fig. 2.2: Four different waveforms and their spectrograms. Example waveforms and spectrograms of four voiced consonants articulated by one of the speakers. Alveolar consonants /da/, /za/, /na/ are, in order, plosive (with long negative VOT), fricative and nasal sounds according to their manner of articulation. Another example of nasal sounds is /ma/, but it is a bilabial consonant, not an alveolar.

The participants moved on to the actual experiment after answering all seven questions correctly. Each listener received 80 trials in one session, in a total of 5 sessions. Trials were randomized so that, when a speaker is picked once, the same speaker has a higher probability $(p = 0.975)$ of being chosen again. We chose this procedure to have a more natural listening setting for the participants. A trial consisted of 1 second of silence, followed first by 400ms of fading out white noise, and then by 200ms of silence, and then by a 400ms-long noise masked syllable. There were 2.1 seconds of extra silence after every 5 trials, and 3 minutes of break inserted after every session. At each trial, after hearing the stimulus, the listener had to respond by selecting a consonant among 16 pre-trained labels that were randomly placed on a 4*4 button matrix for each participant, but that participant-tailored placement of the labels was kept the same throughout the experiment. A new trial started when the participant made a choice.

2.3. Perceptual distances

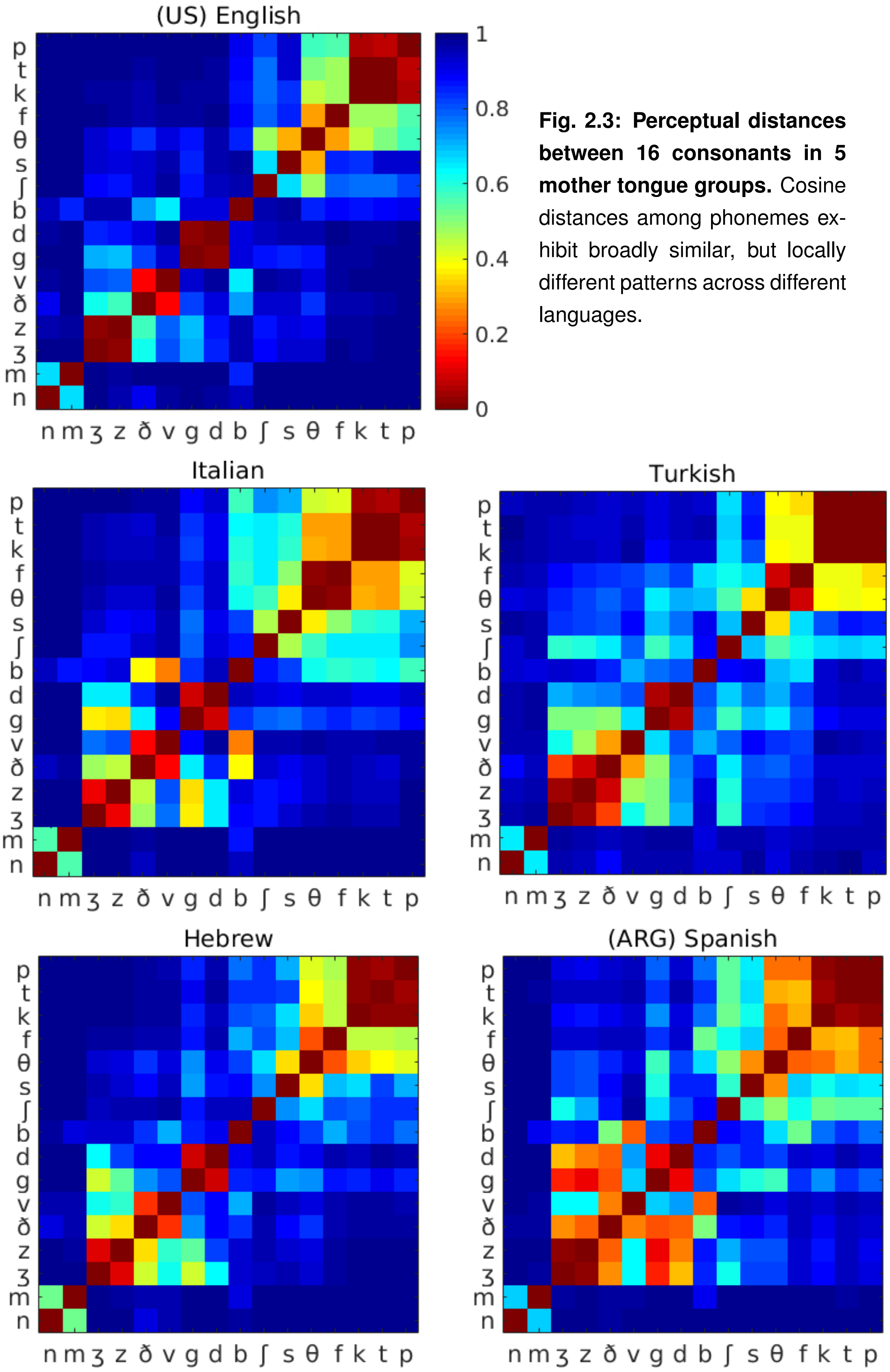
For each language group, the confusion occurrences between the 16 phonemes under consideration obtained from each participant is be pooled together and organized in a 16-by-16 confusion matrix C. We denote its rows by $C_1, C_2, ..., C_{16}$, where the j-th entry of C_i represents the confusion frequency between the $i-th$ stimulus and the $j-th$ response. From this raw data, we can derive the pairwise perceptual distances among the vectors C_i and C_j. One can use different metrics to estimate the distances, here we choose to represent the data as it is, and we use cosine distance metric, D_{ij}, which is defined as:

$$D_{ij} = 1 - cosine\_similarity(C_i, C_j) \tag{2.1}$$

where the cosine similarity between two vectors is given by $cos(\theta)$ that is the cosine of the angle between them:

$$cosine\_similarity(C_i, C_j) = cos(\theta) = \frac{C_i C_j}{\sqrt{(C_i C_i)(C_j C_j)}} \tag{2.2}$$

The calculated cosine distances result in a symmetric matrix that is bounded in $[0, 1]$ (because the C_{ij} entries are all positive frequencies) as shown in Fig. 2.3. On the perceptual distance matrices, phonemes that are perceived as closer to each other form block structures, which given the



(US) English

Italian

Turkish

Hebrew

(ARG) Spanish

Fig. 2.3: Perceptual distances between 16 consonants in 5 mother tongue groups. Cosine distances among phonemes exhibit broadly similar, but locally different patterns across different languages.

ordering chosen are seen mainly around the diagonal. For example, the block formed by the phonemes /p/, /t/, /k/, /f/, /θ/ is easily recognizable in the perceptual phoneme representation of every mother tongue. Although the same block structures are observed for every language group, indicating similar trends in confusions, the degree of similarity within a block varies from one language to another. We will investigate the nature of these patterns to understand the structure underlying the consonant space and its cross-linguistic organization.

A first attempt to quantify the difference between the perceptual spaces of different languages would be comparing the cumulative distribution of matrix elements at different distance values, in each perceptual matrices, as shown in Fig. 2.4. This establishes a ranking between the performance of language groups which might reflect the quality of English language education for the four mother tongue groups other than the native speakers. For distance values that range between 0.9. and 1, English speakers have already the majority of their matrix elements, which shows that they can perceive the differences between their native phonemes well. Indeed for any threshold of dis-

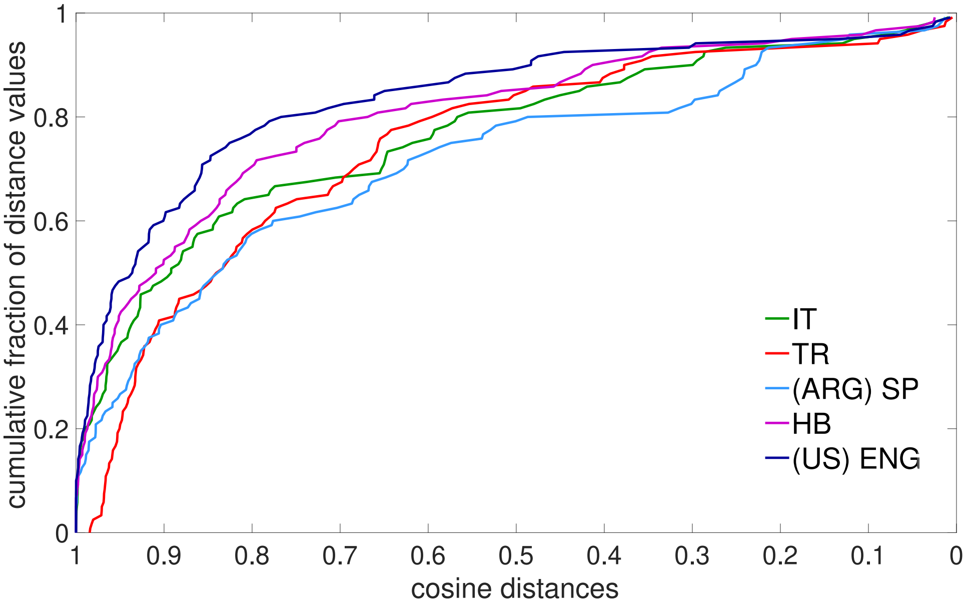


Fig. 2.4: Performance of 5 mother tongue groups as reflected in the cumulative distribution of distances between phonemes. The majority of native English speakers' matrix elements show a value between 0.9 and 1. Hebrew speakers follow their performance closely, whereas Spanish speakers have the majority of their distance matrix elements at values lower than 0.8.

tance, their performance is followed by the performance of Hebrew speakers, while the other three language groups interchange depending on the distance value. Distinguishing the differences between the phonemes is the most difficult for Spanish speakers who have the most of their matrix elements with small distance values. Although understanding the differences concerning the performance is informative, it is not enough to understand the structure of the consonant representation.

2.4. The nature of consonant representation: is it less flexible?

We can visualize the perceptual distances on a two-dimensional space to understand how the emerging structure differs from the square that would represent ideal performance, with no confusion between distinct phonemes. This is similar to the way we represent the vowels (see later chapters); we reduce the high dimensional space the phonemes live in to a plane. Given the more discrete nature of consonants, with their quasi-binary choices: to affricate - or not to affricate, to voice - or not to voice, to labiate - or not to labiate, are they more rigid in their position on the plane than vowels, the position of which change from one language to another?

In Fig. 2.5, Fig. 2.6, Fig. 2.7, and Fig. 2.8, we use the distances to draw a quadrilateral representation, but every quadrilateral is placed on the tile so that its center is aligned to the center of the larger square whose side length is the maximum distance which equals to 1. Therefore, the lengths of the sides of the quadrilaterals do not strictly measure distances, but the quadrilaterals themselves give a visual impression of the common and language-specific distortion of the underlying tile. When considering a larger group of phonemes or phoneme categories, the deformation can also be revealed by making use of an algorithm like a self-organizing map, as we do with the perception of vowels in the next chapter.

2.4.1. Planes of manners of articulation

First, we have a look at what one may call manner planes, where the two dimensions are affrication and voicing. In Fig. 2.5, we use five confusion matrices of Miller and Nicely at different SNR levels, to illustrate how the manner plane gets deformed and squeezed as the noise level increases.

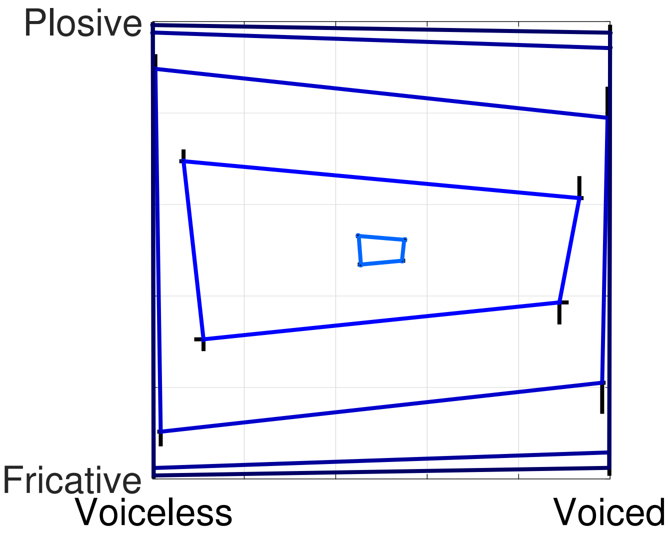


Fig. 2.5: Adaptation of the manner plane (Miller and Nicely data). The manner plane becomes closer to a square metric as the SNR level increases. Here the quadrilaterals respectively correspond to the confusion matrices of SNR:$-18dB$ (the innermost), $-12dB$, $-6dB$, $0dB$, $6dB$ (the outermost). The error bars indicate the SEM.

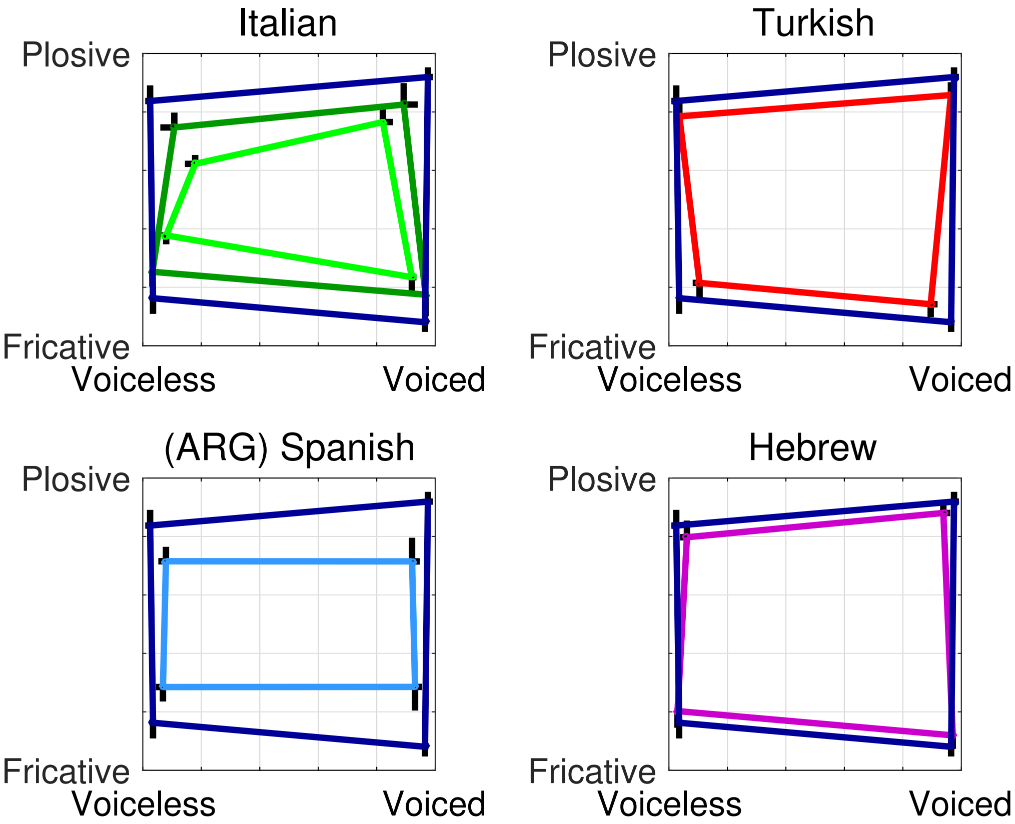


Fig. 2.6: Language-specific remapping of the manner plane. The outer quadrilateral in dark blue represents the manner plane of (US) English speakers, which shows their superiority in distinguishing the two dimensions. The inner quadrilateral in light green in the Italian space is of the confusion matrix for SNR:$-6dB$. Spanish speakers have difficulty in discriminating for affrication regardless of the voicing feature.

The innermost quadrilateral, in blue, corresponds to the perceptual distances in the confusion matrix of SNR:$-18dB$, and as the SNR level increases, the representation obtains almost the form of a square metric as it is illustrated with the outermost quadrilateral, in black, which corresponds to the perceptual distances in the confusion matrix of SNR:$6dB$.

We visualize our data the same way in Fig. 2.6. Within each manner plane, we represent the one of native English speakers in dark blue, which surrounds the planes of other language groups as a result of native speakers' better performance, on average, in discriminating their mother tongue phonemes, as discussed in the previous Section 2.3. If compared to the planes of Miller and Nicely data, the plane of English speakers we tested at SNR:$0dB$ resembles more the plane of English speakers tested at SNR: -6dB by Miller and Nicely as illustrated in Fig. 2.5. However, we should note that, perhaps due to some differences between the stimuli used in the two experiments, while the perceptual plane of our data of native English speakers has a shorter edge between the groups of sounds belong to voiceless fricatives and voiceless plosives, the plane of the data of Miller and Nicely at SNR: $-6dB$ is reshaped in a way that the opposite edge, i.e the edge between the voiced fricative and voiced plosive consonants, is shorter.

On the other hand, comparing it to the four other mother tongues tested at the same SNR level, we observe a close match between the spaces of English and Hebrew speakers as there is also in their performance (see Section 2.3). For the plane of Italian speakers, the inner quadrilateral, in light green, shows the confusion matrix obtained at SNR:$-6dB$. The similar deformation of the Italian plane at two different SNR levels indicates a closer representation of voiceless plosives and fricatives than their voiced counterparts, which is also true for their perception of voiceless and voiced plosives. Argentinian Spanish speakers can distinguish between voiced and voiceless sounds, but for them the difference between plosives and fricatives is diminished compared to the four other language groups, hence there is the narrowing of the affrication dimension.

2.4.2. Planes of places of articulation

The shorter fricative-plosive distance in Argentinian Spanish runs against the idea that consonants preserve their place in perceptual space better than vowels do. It seems that for different languages, different manners might distort the consonant space in specific ways. For example, in Mandarin Chinese the phonetic contrast is on the aspiration dimension, and not on the voicing, which would probably result in a taller than wider basic square, that is the opposite of what we see in the Spanish space. What about the place of articulation? Is the place of articulation a more robust feature in terms of categorical perception, and less subject to deformation across languages?

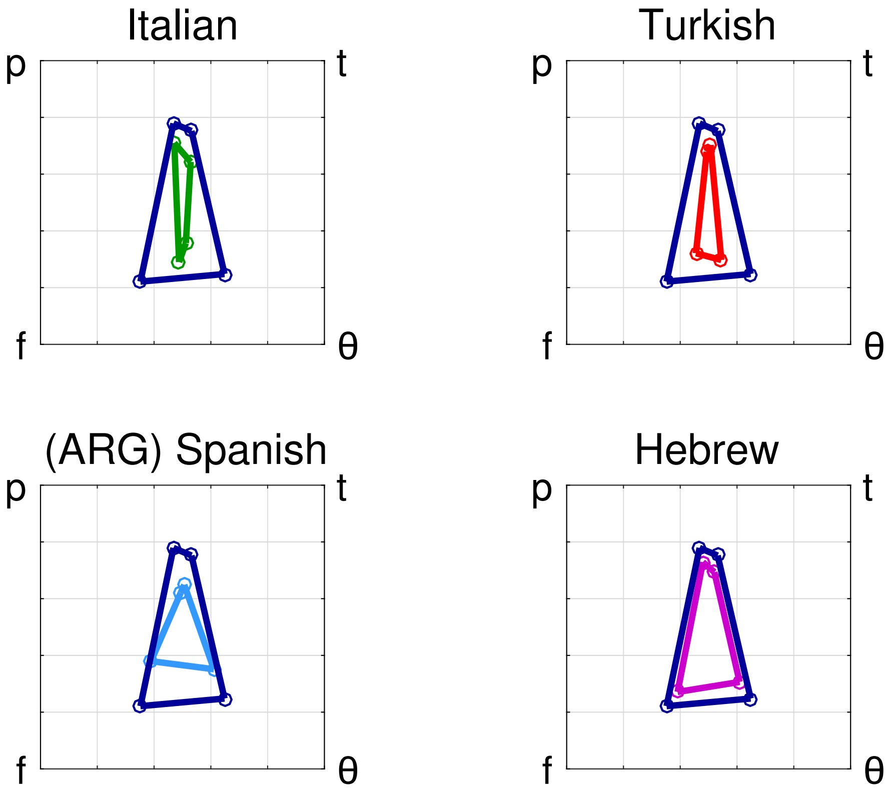


Fig. 2.7: Language-specific remapping of the (voiceless) place plane. Keeping the voicing feature the same, the distance between labial and coronal consonants changes depending on their affrication. Except for Italian, voiceless plosive sounds are perceived closer than voiceless fricative sounds of approximately the same distance. The outer quadrilateral in dark blue represents the perception (also considerably distorted, at this SNR level), of (US) English speakers.

In Fig. 2.7, we look at the deformation of the plane defined by four voiceless consonants, where the dimensions are the affrication and the place of articulation. High confusion frequencies give rise to extremely short distances between different places of articulation, here labial and coronal.

The deformation in every language other than Italian displays a similar structure, whereby the perceptual distance between /p/ and /t/ shrinks compared to the perceptual distance of the fricative pair with approximately the same place of articulation contrast. In the Italian perceptual space, however, the pattern is reversed, hence /f/ and /θ/ (which is not used as a contrasting phoneme from /t/) are represented closer than their plosive counterparts.

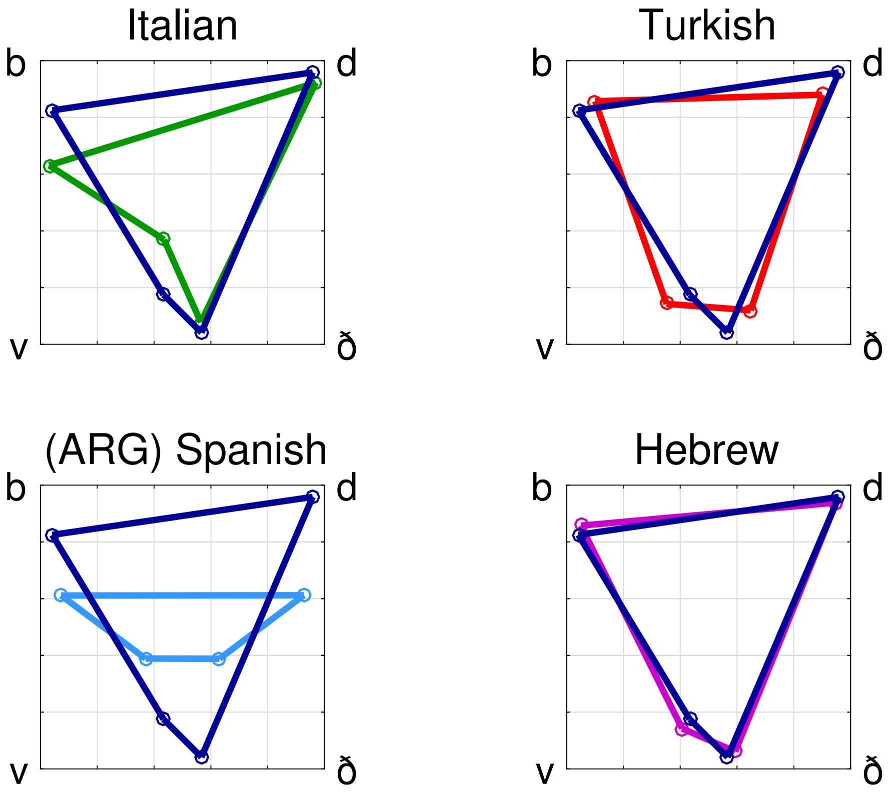


Fig. 2.8: Language-specific remapping of the (voiced) place plane. For voiced consonants, the perceptual distance between labial and coronal consonants changes depending on affrication in reverse fashion to their voiceless counterparts (see Fig. 2.7). Voiced fricative sounds are perceived closer than voiced plosive sounds with approximately the same place of articulation contrast. The outer quadrilateral in dark blue represents the plane of (US) English speakers.

The same finding does not replicate when we look at the same plane, but now for the voiced sounds, as shown in Fig. 2.8. The voiced fricative sounds are perceived closer than the voiced plosive sounds with approximately the same contrast. The distortion of the Spanish space stands out, but that is because their discrimination between /b/ and /v/, and between /d/ and /ð/ is poorer. The loss of the phonetic contrast between /b/ and /v/ is a phenomenon called betacism that is well known to have taken place in Spanish [Macpherson, 1975]. The perceptual distance between /b/ and /v/ is also short for native Italian speakers, but that is not the case for their perception of the pair /d/ and /ð/, and therefore results in an asymmetric distortion.

2.5. The structure of consonant perception

Phoneme representation in the brain has been long studied with the notion of a tree-like structure of hierarchically clustered discrete items [Bouchard et al., 2013]. While this is especially dubious for vowels, which are articulated largely by acting differentially on two simple analogue dimensions within a continuous space, consonants are subject to a complex mixture of analog and discrete operations, which is also supported by the deformation of different perceptual planes in different ways and to varying degrees as discussed in the previous Section 2.4. As we have seen, manner planes, for the majority of the language groups but not for all, maintain a square-like form that implies well-separated perceptual dimensions, while place planes are more distorted and do not preserve the ideal shape. Therefore it remains to assess the adequacy of imposing consonant perception on structures like dendrograms. Is there a hierarchical structure underlying consonant perception? Furthermore, do different languages use similar structures?

2.5.1. Metric content

Let us consider a generic experiment which yields the analog of a confusion matrix C. Then we can extract the frequency of correct performance (that is, the trace of the confusion matrix, i.e., f_{cor}), as well as the mutual information:

$$I_{ml} = \sum_{i,j} C_{i,j} log_2 \frac{C_{i,j}}{P(i)C_j} \tag{2.3}$$

where $P(i)$ is the a priori frequency of each stimulus category, i.e., $P = 1/16$ in our experiment.

Whether the original set of stimuli or external correlates is discrete, e.g. a set of faces presented to a monkey [Treves, 1997], or continuous, e.g. a set of rodent positions in the environment [Stella et al., 2013], it is discretized for the purpose of the analysis, following which a metric content index can be used to quantify the amount of the structure embedded in the perception.

The theoretical minimum of information given f_{cor} is when all "wrong" responses are evenly distributed:

$$I_{min} = log_2 S + f_{cor} log_2 f_{cor} + (1 - f_{cor}) log_2 [(1 - f_{cor})/(S-1)] \tag{2.4}$$

In this case of phonemes, this occurs when the listener does not recognize any further similarity besides the ones already existing in the same phoneme category. As the probability of making a wrong choice is equal for all other responses, perception may be represented by a one level tree with a single mother node, where choosing any one of the (wrong) child nodes is equally probable. In contrast, the information is at the maximum limit when all wrong responses for each phoneme category are concentrated in a single super-category of size $1/f_{cor}$, and are all chosen randomly (including the correct response) within that category:

$$I_{max} = log_2 S + log_2 f_{cor} \tag{2.5}$$

The errors in this case fall into the correct category, but the element or category within the super-category is chosen with equal probability, suggesting a tree-like hierarchical structure underlying perception. All the intermediate cases between these two limits can be quantified by the metric index λ, which is a measure of the concentration of errors and is defined as:

$$\lambda = \frac{I - I_{min}}{I_{max} - I_{min}} \tag{2.6}$$

as visualized by a leaf diagram in Fig. 2.9 where the blue and red lines respectively stand for the cases of minimum and maximum information.

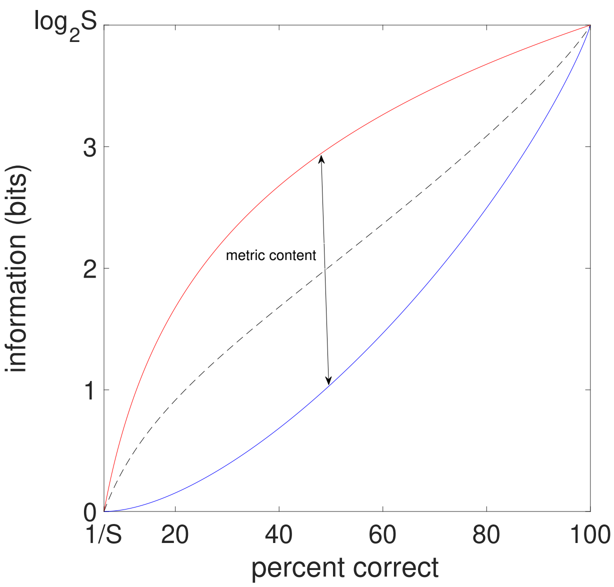


Fig. 2.9: The I vs. f_{cor} “leaf" diagram and the metric content index. As adapted from [Ciaramelli et al., 2006]. Blue, (dashed) black, and red lines corresponding respectively to $\lambda = 0, 0.5, 1$. S is the number of elements in the stimulus set, in this case $S = 16$.

The analyses of Miller and Nicely have focused on the two measures, mutual information and the percent correct score, separately, as did the various other studies that followed their experimental and analytical approach to understand the articulatory dimensions that serve to discriminate the different consonants [Miller and Nicely, 1954, Cutler et al., 2004, Cooke et al., 2010]. However, these two measures are not completely independent as, I largely covaries with f_{cor}, since the former quantifies the concentration of responses across categories, and the latter their concentration in the correct category. The metric content index, λ, we use in this analysis defines the range of values I can take for a fixed f_{cor}, hence transforming I into a measure of the concentration of errors only, and informing us on the perceived structure of consonant representation. It is high whenever responses are systematically clustered into (correct or wrong) distinct categories, i.e. there is a concentrated distribution of errors. For example, if the listeners constantly confuse dental fricatives (/θ/, /ð/) with two alveolar plosives (/t/, /d/), due to their close places of articulation, the corresponding metric content index will be high. In another high λ example, the listeners might perceive more refined similarities and confuse only fricatives among each other, such as confusing the post-alveolar fricatives /s/ and /ʃ/ with the nearby alveolar fricatives /s/ and /z/ on the vocal tract, but not with further away dental fricatives /θ/ and /ð/. Metric content, instead, is low when such relationships of being close or distant are not relevant for perceptual decisions, and in a third example, the listeners confuse the dental fricatives with labial, alveolar, or post-alveolar fricatives equally likely.

2.5.2. A propensity to perceive structures?

Fig. 2.9 illustrates a tree-like representation of the metric content index of perceptual confusions under different categorical constructions at each level. The top panel of the figure shows this quantity measured from the perceptual decision among 16 phonemes (with no forcing into made-up phoneme clusters) by speakers of 5 mother tongue groups, together with the data of native English speakers tested by Miller and Nicely at different SNR levels. What is interesting is that almost all languages fall on the metric content index $\lambda = 0.5$ line or thereabout, despite the differences in their performance of correct identification. This tells us that these 16 sounds themselves do not constitute single isolated categories, but rather there is some clustering structure which encompasses them.

How do the sub-phonemic features shape the perceptual decisions? We can decompose the perceptual space, and categorize the subject responses according to the sub-phonemic features; manner or place of articulation as shown in the middle panel of Fig. 2.9. When grouped into 5 manner categories, that are voiceless plosives, voiced plosives, voiceless fricatives, voiced fricatives, and nasals, the metric content index approaches towards the minimum limit for many mother tongue groups, especially for the confusions of Turkish speakers and for the confusions of English speakers obtained at high SNR levels, but not for Spanish speakers who conserve their place around the 0.5 line (see the left plot of the middle panel in Fig. 2.9). The metric content index of the same data points are even closer to the lower edge of the leaf diagram if the confusions are grouped into six different places of articulation that are bilabial, labiodental, dental, alveolar, post-alveolar, and velar (see the right plot of the middle panel in Fig. 2.9). While five different manners seem to have a further categorization, the sub-phonemic feature place of articulation seems closer to be simply categorical.

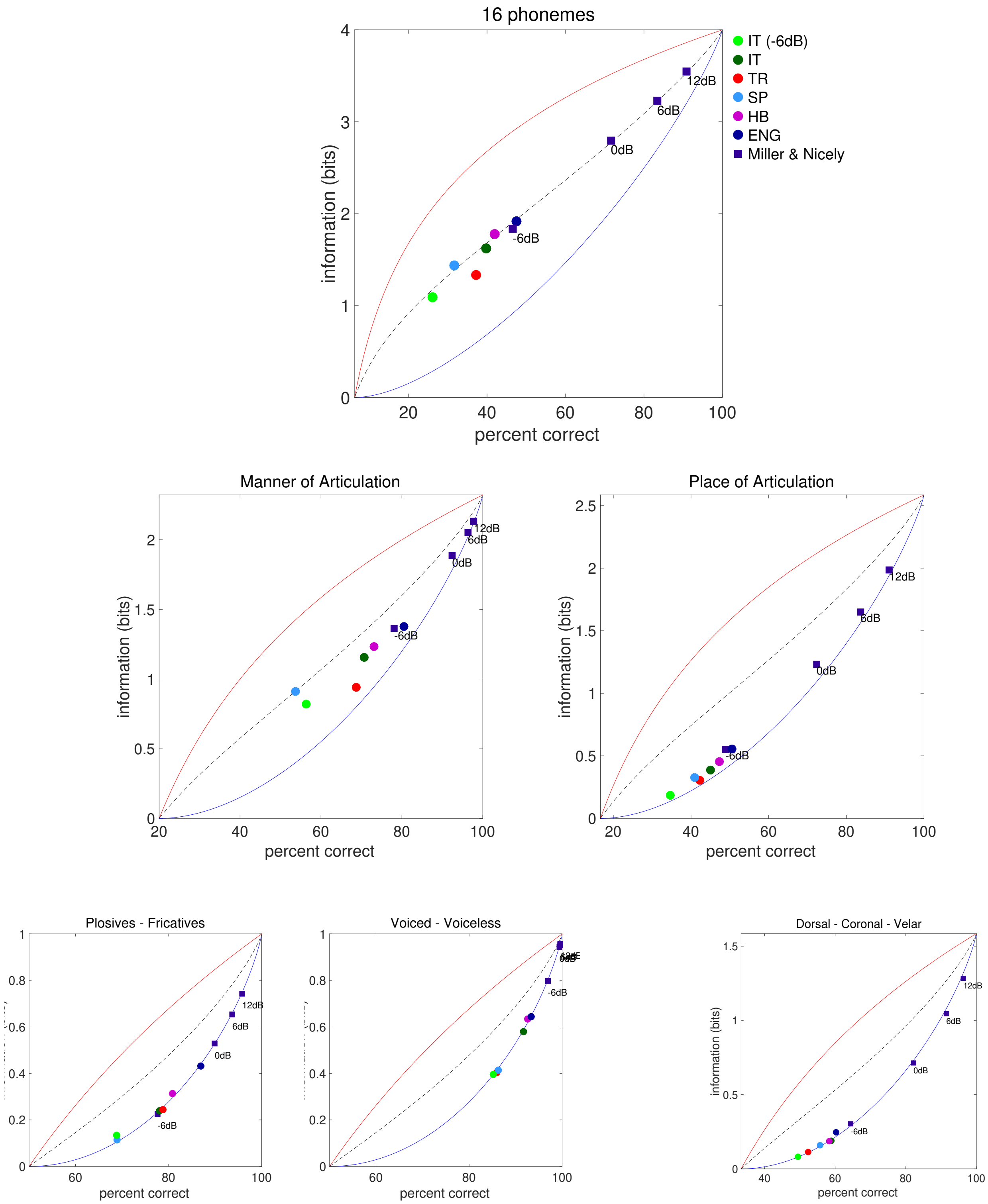


Fig. 2.10: The metric content tree of consonant perception. (Top) The metric content index of perceptual confusions among 16 phonemes. (Middle) The same quantity when the perceptual confusions are categorized into groups of 5 manner categories or 6 places of articulation. (Bottom) As expected the minimum limit is reached when the decisions are categorized into gross clusters of manner or place of articulation.

The metric content index drops to zero, as mathematically required, when five manners are further clustered into only two grand sets of manners either by the affrication property (including nasals as fricatives) or by the voicing property. This is because $\lambda = 0$ when there are only two options to choose from (see the two child plots at the bottom under the Manner of Articulation plot of the middle Fig. 2.9). However, one can notice that between two different manners, voicing is more informative about the identity of the phoneme. Similarly, but less obviously, the six places of articulation can be grouped under three large portions of the vocal tract, and then the perceptual decisions fit into a one level tree structure (see the only child plot at the bottom under the Place of Articulation plot of the middle in the same Fig. 2.9).

Following this analysis, a tentative conclusion is that native speakers of different languages perceive a similar amount of structure ($\lambda = 0.5$) in this set of English consonants, even though their performance in discriminating them varies considerably. Especially when there is a high level of noise, it is as if they seek to perceive a structure which is not necessarily there due to the prevailing difficulty of identifying the phonemes.

Such structure is largely in the multiplexing of 3 discrete features: frication, voicing and place of articulation. Even though place of articulation is not a binary variable, and could in principle reflect metric relations between places of articulation along the vocal tract, the fact that its metric content reduces to zero (bottom right plot of 2.9) shows that such metric relations are not preserved in perception. Consonants that are articulated closer to each other are not confused more than those articulated further away in the vocal tract. Distinct places of articulation are treated therefore, at the perceptual end, as essentially categorical variables, also by non-native speakers who confuse them much more than native speakers. When the processing is bimodal, that is when visual cues are available to the listener, it is possible that the place categories might organize more hierarchically [Files et al., 2015]. The well-known McGurk effect demonstrates that the visual information can have a great influence on the perception of consonants [McGurk and MacDonald, 1976].

2.6. Perceptual attractors

An intriguing aspect of phoneme processing is the analogue to digital conversion that sound perception undergoes and how this conversion is influenced by one's own phoneme repertoire. As we show in Section 2.4.2, with poor discrimination by native Spanish speakers among /b/ and /v/ consonants, speakers have often difficulty in discriminating many consonant contrast that are not contrasting in their native languages. Another very well-known case of failed consonant distinction is between /r/ and /l/ by Japanese speakers, who have only one liquid consonant /r/ [Iverson et al., 2003].

One other example of poor non-native phonetic contrast among 16 consonants is the frequency of correct identification of the two consonants that only the English language has (among those we included in our study), which are the non-sibilant dental fricatives /θ/ and /ð/. These two consonants are very rare across all languages around the world languages (occurring in only 7.6 % of languages in the UCLA phonological segment inventory database [CLARIN-NL TDS Curator, 2018]) and they are problematic in second language (L2) acquisition [Brannen, 2011, Werker, 1989]. We found that native English speakers are better at correctly identifying these two sounds than native Italian, Turkish, and Spanish speakers. Yet this is not the case with Hebrew speakers. Although it is thought that, in the course of language evolution, these two interdentals of Semitic languages became alveolar /ʃ/ and /z/ in Hebrew [Saenz-Badillos, 2002], Hebrew speakers were more successful at their correct identification than even native English speakers were.

Unlike the case with adults speakers, various researchers have shown that newborn infants are able to distinguish phonetic contrasts that are not in their native languages until they are 6 months old, when they start assimilating to their own mother tongue and show similar difficulties as adults exhibit in the same language environment [Kuhl, 2000]. Patricia Khul's Native Language Magnet (NLM) theory posits that infants discretize analogue sound patterns into a 'sound map' as they attune their representation of sounds to the acoustic properties of phonemes in their native tongue in an unsupervised fashion [Kuhl, 1991]. Once they develop the map, the phonetic category prototypes 'pull' acoustically similar tokens of the same phonetic category like a *perceptual magnet*, while non-prototypical members of the category do not have the magnet effect, nor do non-native phonemes. For example, /t/ wraps perceptual representation of /θ/ in Turkish towards itself [Yildiz, 2005] (so do the previously mentioned /r/ in Japanese and /b/ in Spanish).

The adaptation of the perceptual space through experience induces *categorical perception effects* – an increased sensitivity at category boundaries. In another seminal paper in the literature of phoneme perception, Griffith and colleagues have shown that along the /b/-/d/-/g/ continuum created, the subjects had better discrimination at sharply defined phoneme boundaries, where they assign sounds into different categories, than they had in the middle of a phoneme category, where they normally put two sounds (which are at the same phonetic distance, which is defined by number of equal steps of the starting point of the transition of the second formant to the vowel /e/, as those across a phonological boundary) into the same category [Liberman et al., 1957].

The categorical identification of consonants was replicated by Pisoni and Tash (1974) using a continuum of voice-onset times (see the first chapter for its definition). They observed that the participants showed longer reaction times at the category boundaries [Pisoni and Tash, 1974]. Indeed, categorizing an ambiguous stimulus in a longer period is consistent with the pattern completion process seen in attractor networks in which phoneme categories are given by the stable states of network activity [McMurray and Spivey, 2000]. Such states are called attractors, and they develop over time through Hebbian learning of correlated patterns of synaptic efficacies; so that they 'pull' the representation of other input stimuli toward themselves through network dynamics [Amit, 1989]. The attractor model of Vallabha & McClelland was able to replicate the experimental results of a study in which Japanese adults acquired the distinction between /l/ and /r/ after training with feedback [Vallabha and McClelland, 2007].

Presumably categorical perception is shaped by the memory templates set up by the mother tongue, and thus different attractors exist in different languages. Given our data, can we access the consonant attractors in the 5 different languages we studied?

2.6.1. Looking for consonant attractors

To our knowledge, although the seminal study of Miller and Nicely has been replicated several times with different native and nonnative language groups, it has never been studied in the attractor framework, but only through the analysis of the confusion matrix, treating the categories as given in terms of distinct sub-phonemic features [Cutler et al., 2004, Cooke et al., 2010].

As we did for the analysis of the metric content index in Section 2.5.1, we can again construct a 5x5 (for 5 different manners) or a 6x6 (for 6 different places) confusion matrix by grouping all the sounds in a category into one and re-tabulating the frequency of responses among these larger categories of sounds. From these smaller confusion matrices, we can subtract common effects, and obtain relative matrices of each language compared to the four other mother tongue groups:

$$relative\_C_{x_{i,j}} = C_{x_{i,j}} - \frac{1}{5}\sum_{l=1}^{5} C_{l_{i,j}} \tag{2.7}$$

Where C_x is the raw confusion matrix of the mother tongue group x. Accordingly, the sum of the relative confusion frequencies of all mother tongue groups at each cell *i,j* would be 0. In Fig. 2.11, we show the relative response frequencies arranged into 5 manner categories, and in Fig. 2.12 we show the same organized according to 6 places of articulation.

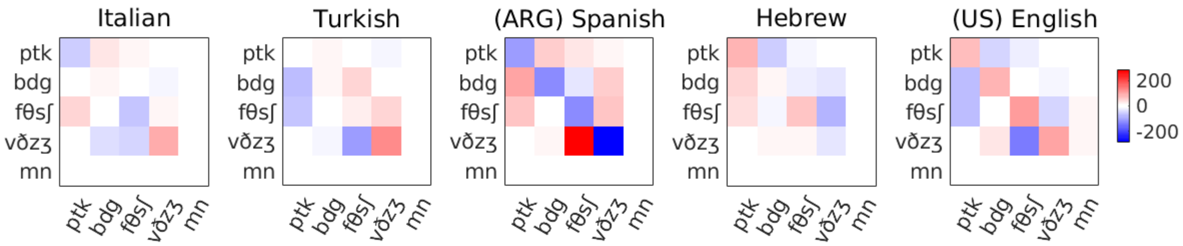


Fig. 2.11: Relative response frequencies tabulated across manner categories. American English participants show stronger (correct) perception along the diagonal, as expected. Hebrew speakers are close to their performance. Spanish speakers are worse at identifying manner categories. They tend to interpret friction as noise, particularly if voiced.

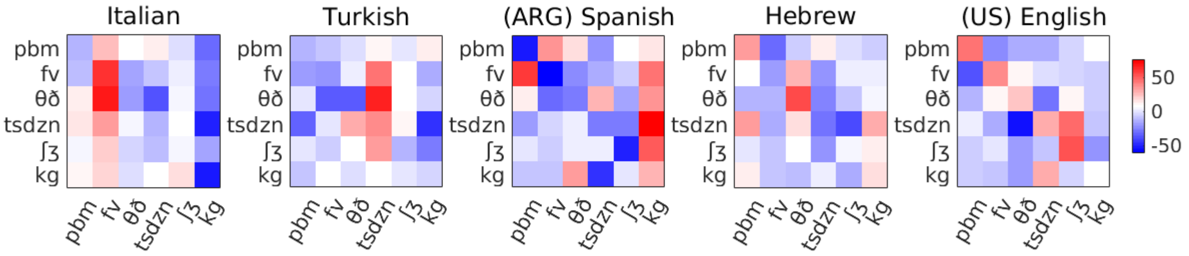


Fig. 2.12: Relative response frequencies tabulated across place categories. Similar to their better performance with manner categories (see Fig. 2.11), native English speakers show a redder diagonal reflecting their better identification of places of articulation. Notice the red columns at different places of articulation for each language.

We then define a simple measure of attraction for each phoneme category m as the difference between the number of times the phoneme category m is presented and the number of times it is chosen as in the response:

$$relative\_attraction_m = \sum_{j=1}^{n} relative\_C_{m,j} - \sum_{i=1}^{n} relative\_C_{i,m} \tag{2.8}$$

For each language, we can plot the relative attraction measure as bars (or in 2D planes); downward ones are those that were chosen more times than they were presented in the experiment. Therefore, the sound categories pulling the perception towards themselves like a magnet, even when they were not presented, are the attractors with negative relative attraction value. On the other hand, upward bars are the consonant categories that were chosen fewer times than they presented the correct answer; the ones that repel the perception away towards other categories.

2.6.1.1. Manner attractors

The relative attraction measure among different manners of articulation across 5 mother tongue groups can be seen in Fig. 2.13. Relative to the other mother tongue groups, Hebrew and Spanish speakers are attracted towards plosive sounds. While Hebrew speakers show a bias towards voiceless plosives (/p/, /t/, /k/), voiced plosives (/b/, /d/, /g/) are attractive for Argentinian speakers. The speakers of the other three groups are biased towards fricative and nasal sounds. Particularly voiced fricatives (/v/, /ð/, /z/, /ʒ/) were chosen more times by Italian and Turkish speakers. Native

English speakers were attracted more towards voiceless fricative consonants(/f/, /θ/, /s/, /ʃ/). Nasal consonants (/m/ and /n/) did not become attractors for any language groups, (except slightly for native English speakers compared to the average), due to the fact that in our experiment these two sounds were clearly distinguished from the rest (see Fig. 2.3) in agreement with the results of Miler and Nicely, who reported nasality to be highly perceptually salient [Miller and Nicely, 1954].

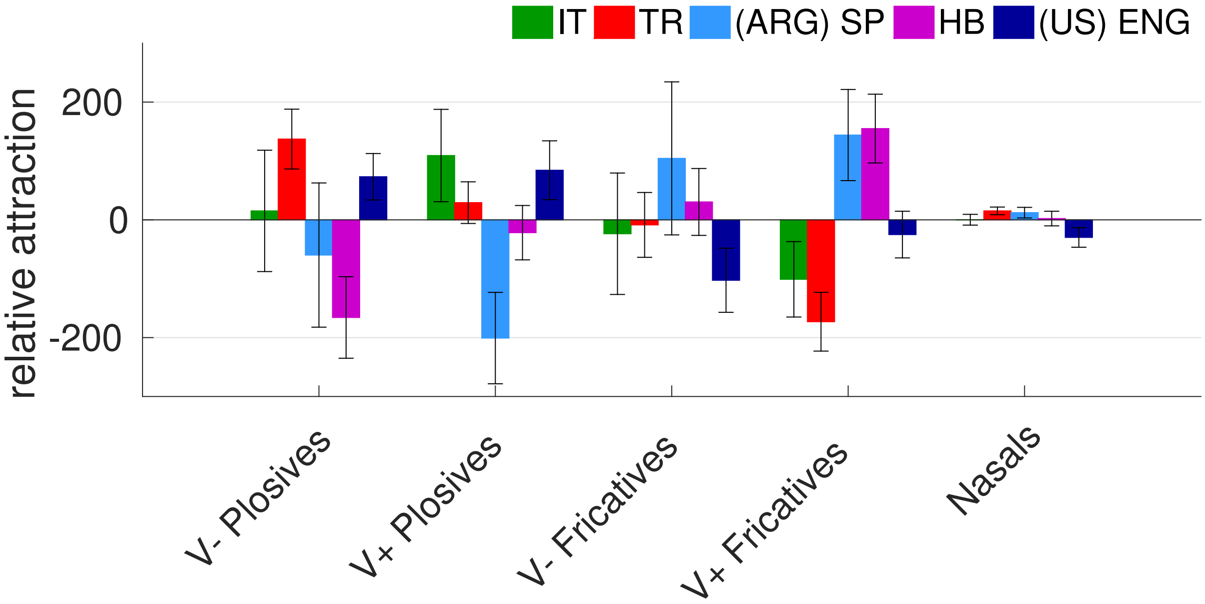


Fig. 2.13: Manner attractors in 5 different mother tongue groups. We plot enhanced perception as downward attraction bars. Relative to the other mother tongue groups, for Spanish and Hebrew speakers, plosive consonants with different voicing properties are more attractive. Italian and Turkish speakers show a bias towards voiced fricatives, while native English speakers are attracted more towards voiceless fricatives. Nasal sounds, as they are well distinguished from the other 14 sounds (see Fig. 2.3), do not exhibit attractive properties except only slightly for native English speakers relative to the rest. Error bars denote the SEM across participants.

If we further group the downward manner attractor bars into two groups of voiced and voiceless consonants, we will see that voiceless sounds are more attractive for Hebrew and English speakers, and voiced sounds are preferred more by Italian, Turkish, and Spanish speakers as response.

2.6.1.2. Place attractors

Overall, for manner of articulation categories, we observe that for some languages, speakers tend to perceive what they hear more as plosive, and for some other languages, speakers are biased to assign affrication property to what they hear. Do speakers of different languages form attractors in different physical locations, where the obstruction most often occurs in the vocal tract, and show a place-specific preference for the sounds they do not clearly hear or do they tend to perceive the consonants with broadly similar places of articulation? Can physical places along the vocal tract be more preferentially attractive across different languages compared to the manner categories?

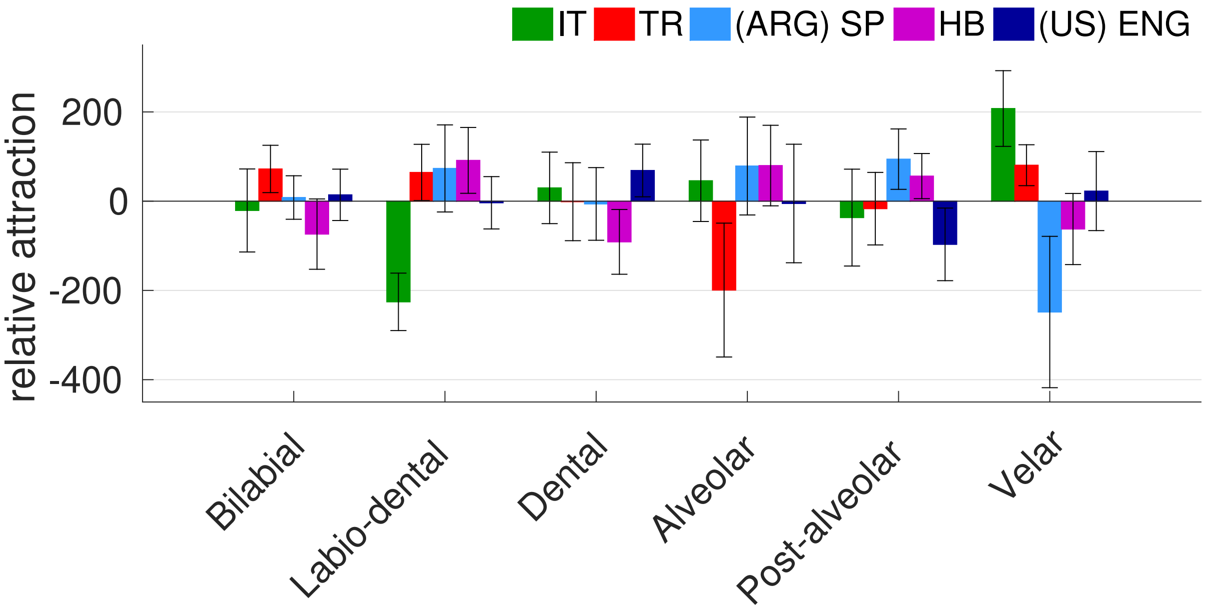


Fig. 2.14: Place attractors in 5 different mother tongue groups. Enhanced perception is plotted as downward attraction bars, as in Fig. 2.13. Relative to the other mother tongue groups, Italian, Turkish, Spanish and English speakers display an attraction towards one specific place of articulation; labio-dental, alveolar, and post-alveolar, and velar sounds respectively. Hebrew speakers show a very weak preference in all three portions of the vocal tract; bilabial sounds in the frontal part of it, dental sounds in the coronal and velar sounds in the dorsal regions. Error bars denote the SEM across participants.

In Fig. 2.14, we see the place-specific attractors across languages. For these five languages, the largest basins of attraction appear to be at different places of articulation. Starting from the frontal part of the vocal tract and going towards the back of it; labio-dental constants (/f/,/v/) are most attractive for Italian speakers. Turkish speakers perceive sounds more as if they are produced at the alveolar ridge (/t/,/s/,/d/,/z/,/n/). Post-alveolar sounds (/ʃ/, /ʒ/) attract the perception of English speakers, while velar sounds (/k/,/g/) attract that of Spanish speakers. Hebrew speakers show a weak preference at particular points in labial, coronal and dorsal parts of the vocal tract. They tend to perceive sounds as one of bilabial, dental, or velar consonants.

Given these language-specific place attractors, the next question we ask is: are there similar tendencies during natural speech perception? Can we then say that place of articulation guides the dynamics of the activity of neuronal populations that implement phoneme perception processing?

2.7. Towards more natural consonant processing

So far we have considered perception of consonant-vowel (CV) syllables, and quantified various difference (and similarities) across different languages. We have designed another experiment with real word stimuli as we wanted to investigate if it is really the case that languages set up stronger attractors at different places of articulation. What would be the reason for it? One reason could be the differences in the frequency of occurrence of the phonemes in every language. After all, what reshapes the representation of speech sound patterns in the infant brain is the exposure to their native language during the first six months [Kuhl, 1991] (Section 2.6). Babies, naturally, lose their discrimination ability of non-native contrasts that they did not experience enough, as their language environment did not contain sufficient instances of them. Therefore, phoneme occurrence frequency might be one of the principal components of phoneme representation throughout later life.

We gathered (or obtained from a written corpus [Oflazer, 2016] in the case of Turkish that has phonologically transparent orthography) phoneme occurrence frequency data from various sources [Hayden, 1950, Tsoi, 2005, Fry, 2004, Blumeyer, 2012, Palmer, 2011, Silber-Varod, 2011, Goslin et al., 2014, Guirao and Garcia Jurado, 1990]. In Fig. 2.15, we plot relative occurrence frequencies of a set of consonants restricted to voiceless sounds laid out according to the place of articulation in the vocal tract; starting from the most frontal, lips, part to the most posterior of it, the glottis. As in Fig. 2.13 and Fig. 2.14, we again plot consonants that have higher frequencies as downward attractor basins.

The attractor basins we previously observed in Fig. 2.14 do not match Fig. 2.15 very well. However, if we reasoned that the mismatch may be due also to the interaction with the voicing dimension, which as we noted in the study of Section 2.6.1.1 plays out differently for different groups of subjects. Therefore, we decided to use a smaller set of consonants than we did in the previous experiment in order to limit the response options, and to make them engage in actual speech perception through word processing. Can we then find attractions towards the different place of articulation at which their mother tongue has higher frequencies, as shown in Fig. 2.15?

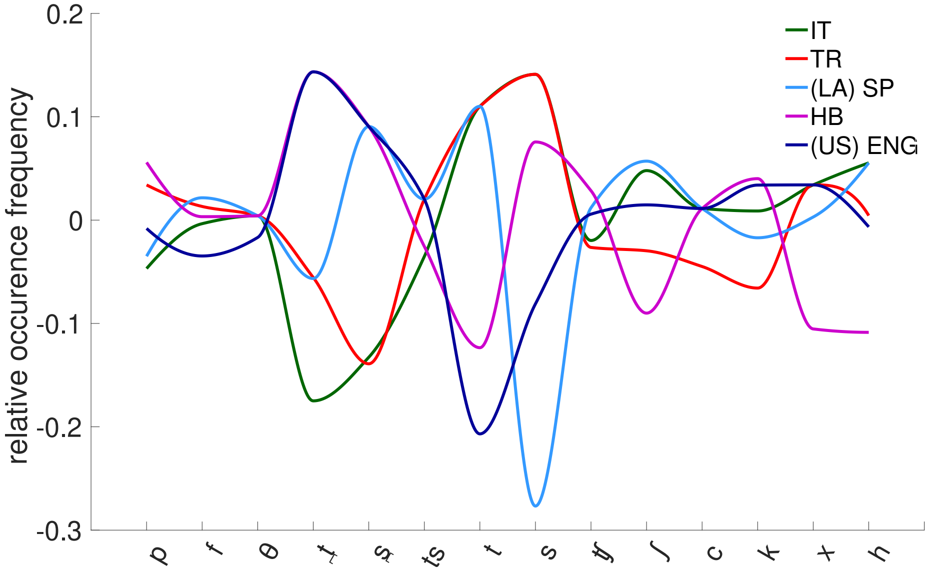


Fig. 2.15: Relative phoneme occurrence frequencies across 5 languages. Different mother tongues vary in their phoneme occurrence frequencies relative to the other languages. Here the x-axis is arranged according to the place of articulation, and consonants with high frequencies are the ones shown with downward basins, indicating attraction towards them. Frequency data is obtained from various sources (see the text).

2.8. Materials and methods

2.8.1. Auditory Stimuli

As in the previous study, we have recorded from five female native American English speakers, but this time articulating 8*3 meaningful (or in a few cases quasi-meaningful) English words, which start with one of 8 English voiceless consonants (/p/, /f/, /θ/, /t/, /s/, /ʃ/ , /k/, /h/) and end with either of /ale/, /in/, and /ore/. All the (8*3*5 = 120) stimuli were first normalized to the same RMS using Praat [Boersma and Weenik, 2018], and then for each word a babble noise mask was added so that the final signal-to-noise ratio would be $-6dB$. The masks were created merging the dialogues of 4 male and 4 female radio artists of American Radio theater.

2.8.2. Participants

Again, we have tested five groups of participants of different mother tongues; Italian, (American) English, Turkish, and (Argentinian) Spanish, and Hebrew. In each language group, there were 12 non-bilingual listeners and all the listeners could speak English.

2.8.3. Paradigm

Each listener went through a training phase before the experiment. In order to be adequately trained, the participants had to listen to the words placed on a 4*6 button matrix by clicking on each one of them. At the first click on each word, the definition of the word would appear on the screen. The participants were instructed to read the descriptions carefully. They could not move on to the experiment phase unless they listened to each word at least two times. During the experiment, each participant received 60 trials in one session, in a total of 6 sessions. Trials were randomized with the same procedure of the previous experiment (see Section 2.2).)A trial consisted of 1 second of silence, followed by the 400ms of fading out babble, and then 200ms of silence, and then the masked word. There was 2.1 seconds of extra silence after every 5 trials, and 3 minutes of break inserted after every session. At each trial, after hearing the stimulus, the listener had to respond by selecting one of the eights words that were randomly placed on a 2*4 button matrix. A new trial started when the participant made a choice.

2.9. Phonemes in the same space

For every language, we have calculated perceptual distances between 8 consonants with the cosine distance metric, the same way we did in Section 2.3. What are the main components that determine the perceptual distances between the phonemes? Are these distances related to physical distances in place of articulation?

For this we can embed the consonants and their pairwise distances onto a lower-dimensional space using a dimension reduction technique like MDS [Torgerson, 1952]. In the new configuration found by MDS, the distances between the consonants on the plane approximates the original distances extracted from the confusion matrix. In order to represent all solutions in one space, we averaged the distance matrices to obtain an "average" low-dimensional embedding. It is important to understand that the embedding in a lower dimensional space is not unique. It is invariant under translation, rotation and reflection. To try to match the space of each mother tongue to the one of the average perceptual distances, we use the procrustes algorithm that rotates each space to reach maximum similarity with the reference space (i.e. the average one) [Kendall, 1989]. Doing so, we exploit the opportunity to observe distances between the same sounds across languages.

2D solutions suggested by MDS analysis for the perceptual confusions among 8 consonants (/p/, /f/, /θ/, /t/, /s/, /ʃ/ , /k/, /h/) in the word initial position, 7 consonants (/p/, /f/, /θ/, /t/, /s/, /ʃ/ , /k/) in the syllable initial position in the previous experiment (see Section 2.2) are shown in the top and bottom panels of Fig. 2.16 respectively. For all 5 languages, we observe 4 similar clusters in both experiments. Alveolar fricative sounds (/ʃ/) are away from the three other clusters. Plosive sounds (/p, /t, /k/) are represented very close to each other together with glottal fricative (/h/) in the experiment with the real words (see top panel). The distances between the plosives are larger when the perception is of syllables, not words. As previously mentioned, among the languages we consider here, /θ/ is unique to the English language, and although in non-English languages, /θ/ is usually taken for /t/, our results shows a closer representation between labio-dental and dental fricatives (/f/, /θ/). Similar to plosives, the /f/ sounds end up closer to each other in word representation than in syllable form, when they lay in between plosives and /θ/. Notice how '/θ/ of English speakers' is slightly separated from /θ/ of other mother tongue groups, and more separated from /f/ when the stimuli is in the word form compared to the other representation of these two in other languages. These fricatives fall in between plosives and alveolar fricative (/s/).

The layout of these consonants on the plane shows groups based mostly on the manner of articulation, and it is similar in both experiments. The clusters, and the sounds within the clusters, especially plosives, are closer to each other, when the perception of phonemes is through processing of single words,

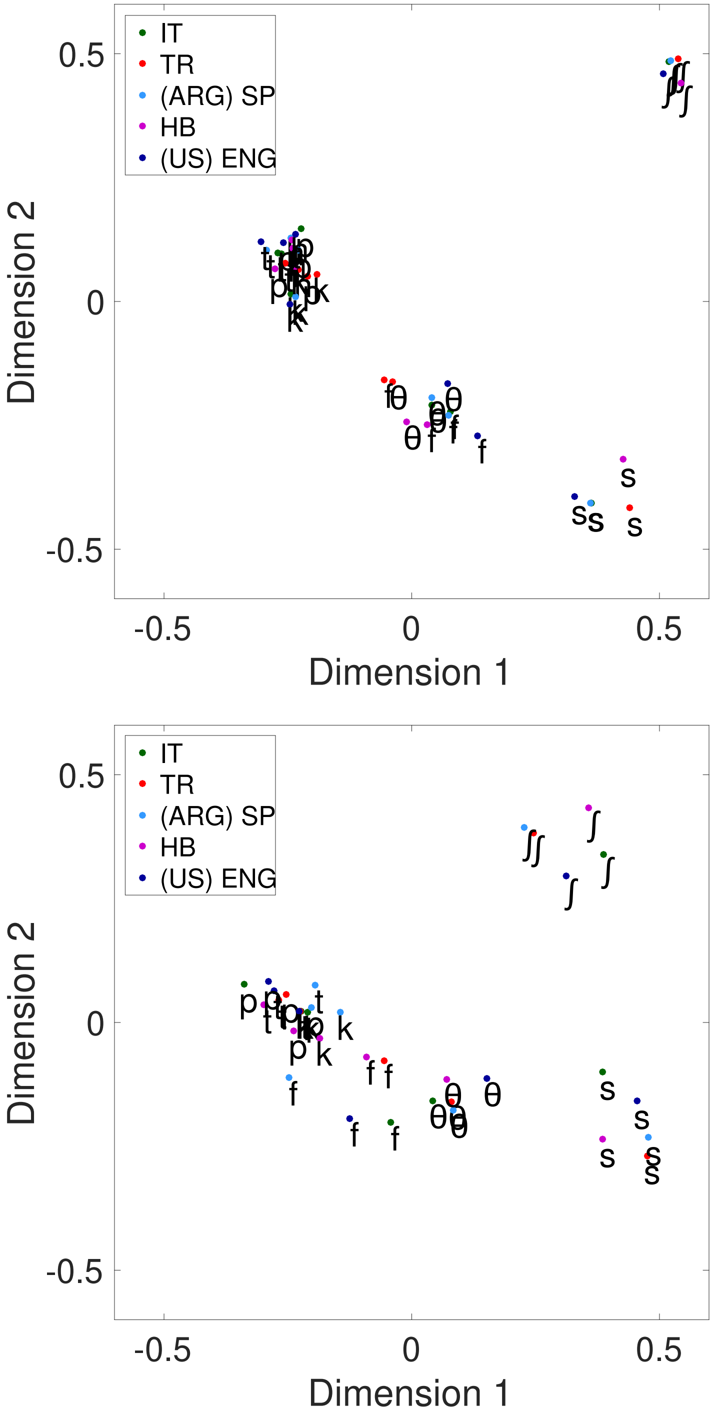


Fig. 2.16: Multi-Dimensional Scaling of perceptual distances between the consonants in two different experiments. For the solutions in both panels, we used the procrustes algorithm to match the lower dimensional space of each mother tongue to the space obtained by the average perceptual distances across languages. This allows us to represent all 5 embeddings in the same space. (Top Panel) Scaled distances of perceptual confusions among 8 consonants in the word initial position. There are 4 clusters of sounds. The fricative post-alveolar /ʃ/ cluster is away from the other three. The voiceless glottal fricative /h/ is grouped together with plosives /p/,/t/,/k/, all of which are very close to each other. The labio-dental and dental fricatives /θ/ and /f/ are in between plosive and the cluster of alveolar fricative /s/. (Bottom Panel) Same as in the top panel, but perceptual confusions are of the data of the previous experiment (see Section 2.2). Here, among 16 consonants we consider 7 that are common in both experiments. Similar clusters are observed as in the top panel, but within the clusters the sounds are less intertmixed.

and not syllables. That is perhaps underlining, once again, the challenge we face when we try to understand each other's phonemes, several of which blend together to make up the words we utter during natural speech.

2.10. Revisiting the consonant attractors

What about attractors? Are speakers of different languages attracted towards sounds that are highly frequent in their mother tongue?

Fig. 2.17 does not support this hypothesis. The attractors suggested in Fig. 2.15 are not in correspondence with the ones we found. Especially the downward bars of Turkish, Spanish, and Hebrew speakers for the alveolar fricative (/θ/) are not in parallel with our idea that there may be a bias for the sounds that are highly frequent in the native language. For example, /ʃ/ are used often by the speakers of Hebrew and Turkish languages, but this alveolar fricative is quite sparsely chosen by the speakers of these mother tongues. Similar mismatches exist for other phonemes.

Admittedly, our analysis is incomplete and biased by a number of factors. For example, while the confusion matrix reflects the limited options available in each particular experiment, in estimating the relative frequency of occurrence of consonants with different places of articulation, in Fig. 2.15, we implicitly considered the entire range of consonants, irrespective of their occurrence or not in a language or in an experiment. Furthermore, in the same figure, we made somewhat arbitrary binary decisions, like considering /t/ as dental in Italian and alveolar in English, which is widely regarded as common but by no means clearcut trend among speakers of the two languages. Given these shortcomings, still we are inclined to conclude that the observed confusion matrices do not appear to reflect in any straightforward manner the frequency of usage of different consonants. We ask, as an alternative hypothesis, whether they might reflect the degree of similarity in the neural signature of the different consonants. Much less is understood, however, about the neuronal representation of consonant phonemes, and virtually nothing about their cross-linguistic variability. We therefore put on hold this latter issue, of the cross-linguistic differences, and take a first step towards assessing differences in the neural response to different phonemes.

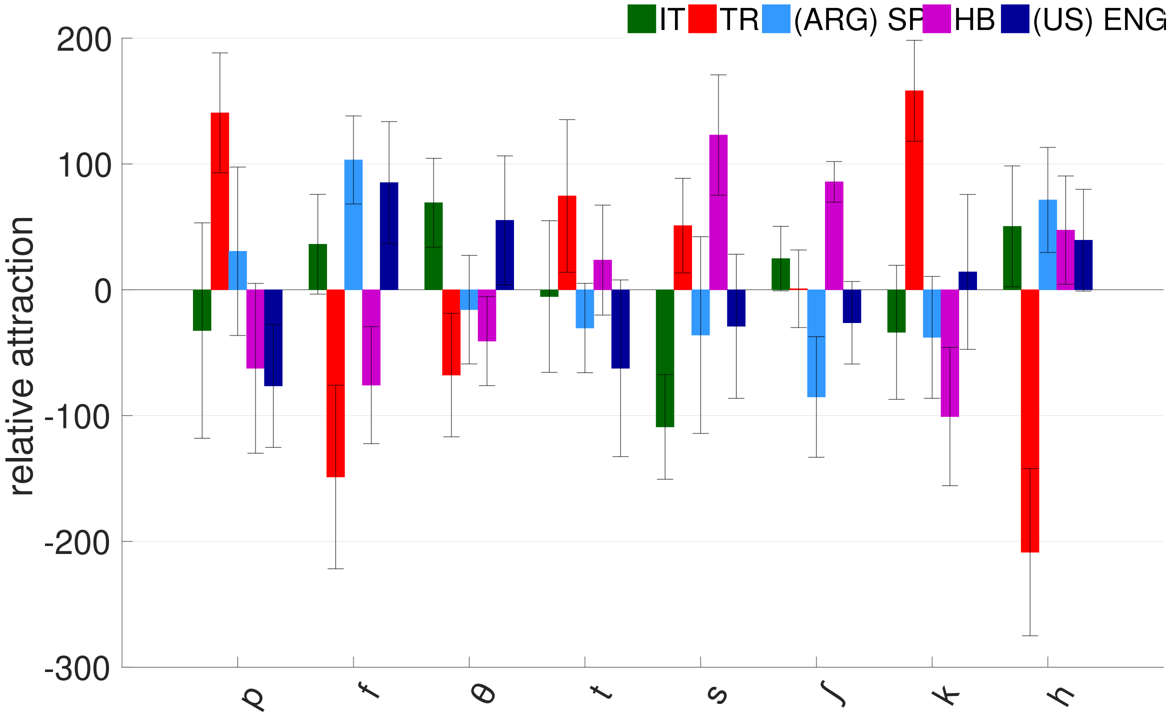


Fig. 2.17: Putative place attractors in 5 different mother tongue groups. Enhanced perception is plotted as downward attraction bars, as in Fig. 2.13 and Fig. 2.14. Speakers of different languages show a preference towards different sounds, but they are not the ones highly frequent in their mother tongue (see the relative phoneme occurences of these 8 consonants in Fig. 2.15). The alveolar fricative (/θ/) is attractive for the speakers of Turkish, Spanish, and Hebrew, but not for native English speakers who have /θ/ in their mother tongue phoneme inventory. Error bars denote the SEM across participants.

2.11. A glimpse into the neural representation of consonants

Here we provide a first look we took for the neural signature of the consonant representation which we initially wanted to study in the oscillation framework [Wang et al., 2012, ten Pever and Sack, 2015]. Later, we were motivated to investigate first the neural representation of vowels (see Chapter 4) and no time was left to complete the study presented here, despite the intriguing preliminary findings described below, which we report as an introduction to the study with vowels.

2.12. Materials and methods

2.12.1. Stimuli

We have recorded eight consonants /p/, /b, /g/, /t/, /f/, /s/, /v/, /z/, followed by one of the four vowels /i/, /a/, /u/, /e/ articulated by a female native English speaker. There were 7 repetitions of 32 consonant-vowel pairs resulting in a total of 224 stimuli.

2.12.2. Participants

There were 11 non-bilingual native Italian speakers participated in the experiment.

2.12.3. Paradigm

During auditory trials, the participants were asked to look at the fixation cross on the screen and pay attention to the auditory stimuli. One auditory trial was comprised of 1.2 seconds of silence followed by 285ms of a randomly chosen auditory stimulus. There were 4 sessions in total, and in every session, each participant listened to all 224 stimuli in a randomized order. We have implemented an attention paradigm to make sure the participants do focus on listening to the phonemes. At 160 random points, one of 80 cartoon like pictures appears on the screen, and the participant is asked to press on the keyboard the first letter of the English word that describes the image. For every phoneme, there are 10 congruent attention trials and 10 incongruent attention trials. In the congruent trials, after the presentation of the phoneme stimulus, a picture of the word that starts with the same sound appears, and in the incongruent trials the upcoming image is of a word that starts with a different sound than the recently heard one. The participants had three breaks of 40 seconds long.

2.12.4. EEG data collection and preprocessing

EEG data were collected in a sound-proof booth. The stimuli were presented at a comfortable and constant volume from headphones. The brain activity was recorded with a 64 channel BioSemi ActiveTwo system (BioSemi Inc., Amsterdam, Netherlands) at a sampling rate of 1024Hz. A Common Mode Sense (CMS) active electrode was used as the reference, and a Driven Right Leg (DRL) passive electrode was used as the ground. Two external electrodes placed on the right and left of the outer canthi, and one external electrode placed under one eye were used to obtain horizontal and vertical electrooculograms (EOG). Two additional electrodes were placed on the left and right mastoids. Individual electrode offsets were kept between ±30 μV. Participants were requested to minimize movement throughout the experiment except when they had a break.

EEG data preprocessing was performed with EEGLAB toolbox [Delorme and Makeig, 2004]. Offline data was imported by reference to the average of the mastoids as common reference averaging is not preferred for studies of auditory evoked potentials [Khalighinejad et al., 2017], and then band-pass filtered (0.1-30Hz). Following the segmentation of the EEG data into approximately 495ms long epochs starting at around 200ms before stimulus onset and 10ms after stimulus offset, bad channels were discarded using the EEGLAB pop_rejchan function [Delorme and Makeig, 2004]. Trials containing extreme values (±200 μV) were eliminated. On average 12% of data was removed for each subject. Independent Component Analysis (ICA) was used to remove eye blinks and muscle artifacts [Delorme and Makeig, 2004, Chaumon et al., 2015]. At this point, the data was divided into the desired conditions, and then it was pruned by randomly discarding trials to ensure the same amount of trials per condition. Finally, missing channels, fewer than 10% of all channels, were interpolated, which was followed by a baseline correction with a reference interval of 200ms before stimulus onset. We normalized the neural response of each EEG channel to ensure zero mean and unit variance for each subject. The resulting dataset of each condition had the data of each participant averaged over trials.

2.13. Do ERP signals reflect consonant features?

Looking at the ERPs of the eight consonants, we observe different patterns in the N1-P2 complex for plosives and fricatives, as shown in Fig. 2.18. The observed negative deflection around 100ms for four fricatives, /f/ /s/ /v/ /z/, is suggestive of voicing processing (see the top right panel of Fig. 2.18). The evoked potentials averaged across all electrodes, show a negative deflection and it is bigger for voiced fricatives, /s/ and /z/ than it is for voiceless fricatives /f/ and /s/. The later component of the ERPs, P200, hints at place processing. Around 200ms, there is larger positive deflection for alveolar consonants than there is for labio-dental consonants. On the other hand, we cannot see the neural signature of voicing or place processing for plosive consonants. This might be explained by their more complex nature, or by the concurrent variation in voice onset times and aspiration, which we naively did not control for, whereas no such parameters are involved in the production of fricatives.

In the bottom panel of Fig. 2.18, we see the averaged ERPs to the same syllables, but this time grouped according to the vowel they end with. The transitioning into four different vowels, /i/ /u/ /e/ /a/, produces the same neural trajectories for both plosives and fricatives. The differentiation of the signal at N100 indicates the modulation of the signal according to the second formant, and hence a difference between the magnitude of the negative deflection for front vowels, /i/ and /e/, vs back, /u/ and /a/, vowels, which is greater for the latter. First formant processing is reflected by the P200 component, which is greater for low vowels, /e/ and /a/, than for high vowels, /i/ and /u/.

Because of our initial motivation to study the neural correlates of consonants, we limited vowels used in the stimuli to these four standard categories. More vowel categories are needed to establish the correspondence between the two components of the ERP signal and the neural processing of vowels according to the first and second formant. The literature on the cortical processing of phonemes in general, but in particular of vowels, is rather confusing. Vowel perception has been mainly studied with the mismatch negativity (MMN) paradigm; it has focused mainly on native vs non-native phonetic contrasts [Näätänen et al., 1999, Dehaene−Lambert, 1997, Näätänen, 2001]. The earlier studies on evoked potentials to isolated vowels (without MMN paradigm) focused only on the N100 component (or it is MEG counterpart N100m), but presented various results on which

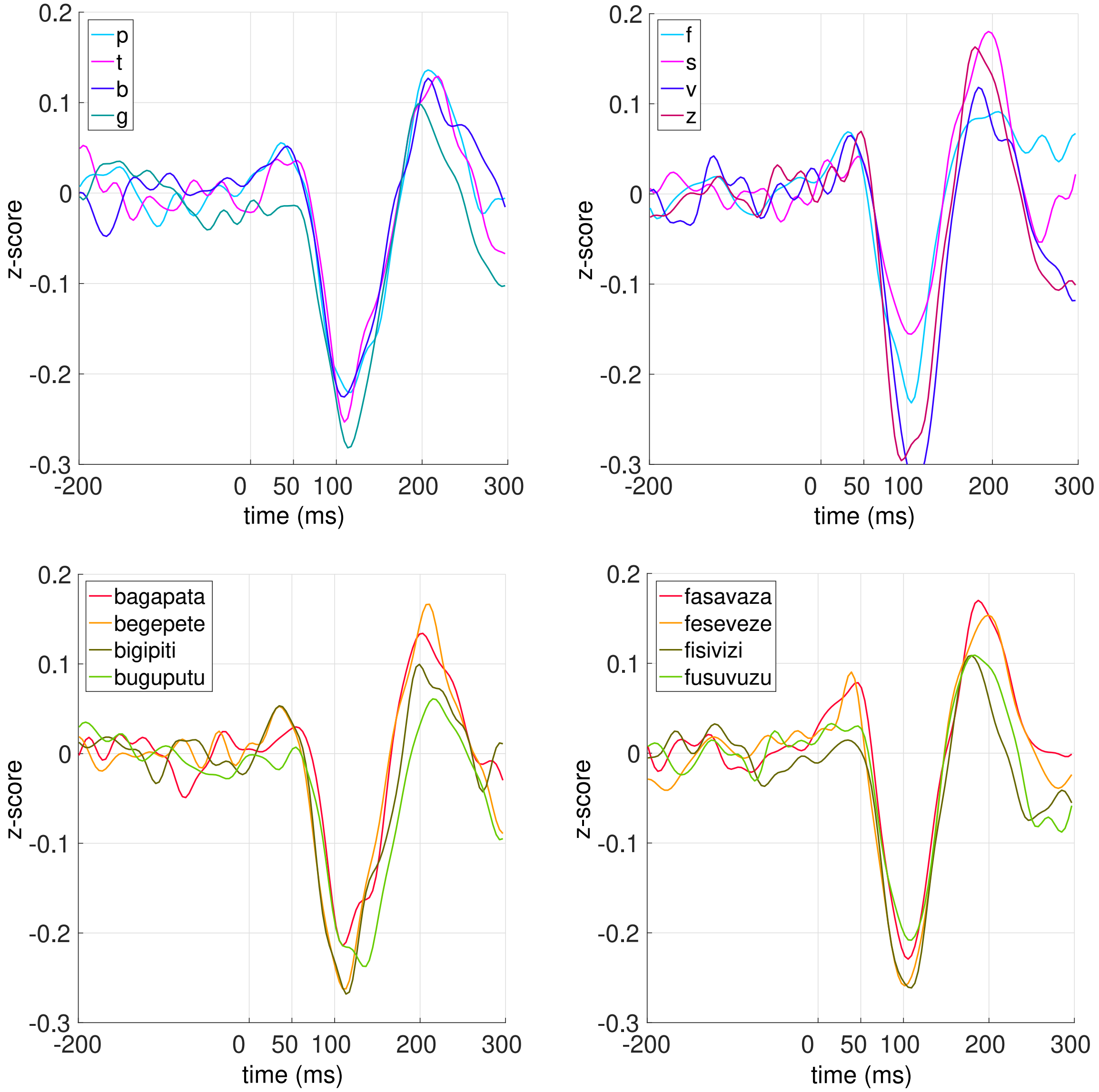


Fig. 2.18: ERPs of CV syllables. (Top) When grouped according to the consonant they start with, either plosive (on the left) or fricative (on the right), we observe, averaged across all electrodes, different patterns. We see a greater negative deflection for voiced fricatives, /v/ and /z/, than for voiceless fricatives, /f/ and /s/. At P200, the neural signal is indicative of place of articulation processing. We did not find such a meaningful clustering of plosives, perhaps due to some aspects, like VOT or aspiration, we did not control for. (Bottom) When ERPs, again across all electrodes, are compared according to the vowel they end with, they show more similar patterns. At N100, we see a greater negative deflection for high vowels, /i/ and /u/, than for low vowels, /e/ and /i/, suggesting processing of the second formant. At P200, a greater deflection for front vowels, /i/ and /e/, than for back vowels, /u/ and /a/, is observed.

there is still no clear consensus. While some studies found the amplitude or latency of N100 to be serving as an index of the degree of opening of the vocal tract (F1) [Grimaldi et al., 2016, Roberts et al., 2000], some others found it to index the place of obstruction (F2) [Obleser et al., 2003] and the peak latency and source localization of it to reflect the difference between the two formants [Obleser et al., 2004].

The rest of this thesis is devoted to the investigations carried out in the more transparent 2D vowel space, motivated by the hope that its transparence will make it easier to gain an insight into the cortical computations of vowel processing (see Fig. 2.19).

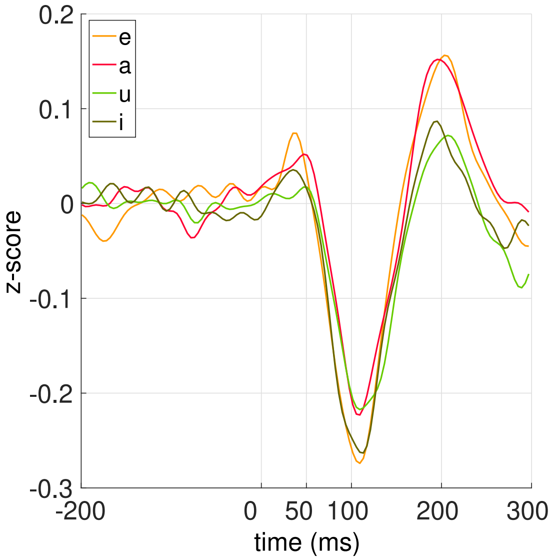


Fig. 2.19: ERPs of CV syllables according to the vowel they end with. The ERPs of all syllables according to one of the four vowels they transition into (same as the bottom panel of Fig. 2.18, but plosives and fricatives are not separated). This modulation of the neural signal by the first (P200) and second formant (N100) that we have found motivates studying the vowel trajectories in the two-dimensional vowel plane.

3

Vowel space: Is it really 2D?

3.1. Rounding: just another dimension?

While consonants are typically brief and are articulated through a complex mixture of analog and discrete operations that result in rather opaque sets of acoustic features, such complexity is conspicuously absent in the case of vowels, which are articulated largely by acting differentially on two simple analogue dimensions: the degree of occlusion of the vocal tract, and the position of the occlusion. As we already explained in Chapter 1, these two variables translate transparently from biomechanics into acoustics, as the first two 'formant' (resonant) frequencies F1 (degree of opening) and F2 (degree of fronting). Therefore 'vowel space' appears simple, and to a first approximation can be characterized by fixed points or short trajectories on a two-dimensional manifold. Some additional acoustic features also reflect in a direct way other analog articulation patterns, such as *rounding* and *duration* (i.e. tense vs lax vowels), which can constitute principal discriminating features for some languages. In this chapter, we focus on the rounding feature which refers to the shape of the lips during vowel production. The vowels that are articulated with protruding lips are called rounded, and the vowels that are produced with spread lips are called unrounded. In contrast to the openness of the vocal tract and the location of the main occlusion that have clear acoustic correlates (namely the F1 and F2 formants), for roundedness it is less clear what the corresponding acoustic correlates are. Therefore vowels have typically been represented on a two dimensional plane with dimensions corresponding to the degree of occlusion of the vocal tract and the position of the occlusion, as can be seen on the left panel of Fig. 3.1. Rounding, on the other hand, does not appear as a separate dimension, but is often summarily embedded into this 2D representation: vowels are treated as single points in the two dimensional plane. The rounded element of the pair is positioned to the right of the unrounded element, as if rounding amounted to a small standard (and time-invariant) shift in the second formant, F2, which would be slightly reduced. An analysis of the acoustic waveform indicates however a different characteristic for rounded vowels: the frequency of their second formant increases in time, whereas for the matching unrounded vowel it appears to be constant or even decreasing. Which is the critical feature, the temporal variation of the second formant, or the slight difference in its temporally averaged value? To address this question, we must first consider whether rounding is a contrastive feature at all for vowel discrimination, in any specific language.

3.1.1. Rounding harmony: a special case for Turkish

Turkish is one of the languages in which rounding is an important perceptual cue. Each one of the eight vowels of Turkish vowel inventory (/a/, /e/[1], /ɯ/, /i/ , /o/, /œ/, /u/, /y/ in their alphabetical order) can be, in Turkish, distinguished from the rest by one or more of three binary features: highness, backness, and roundedness. The vowel /ɯ/ for example, can be realized as emerging from three differential articulatory dimensions: highness positive, frontness negative, and roundedness negative. Such a vowel space defined by this 2x2x2 system can be represented more appropriately by a cube rather than the traditional trapezoid representation, as shown in Fig. 3.1.

Given the phonological rules that Turkish language works on, the aforementioned systematic classification of the vowels is convenient and useful for the speakers of Turkish, which is agglutinative in its nature. One such phonological requirement of Turkish, and some other Turkic languages, is the *harmony* between vowels, in which the vowel in every upcoming suffix must belong to the

[1] Turkish vowels /e/ and /œ/ are realized as mid vowels, whose exact IPA transcriptions are e̞ and ø̞ or œ̝, respectively.

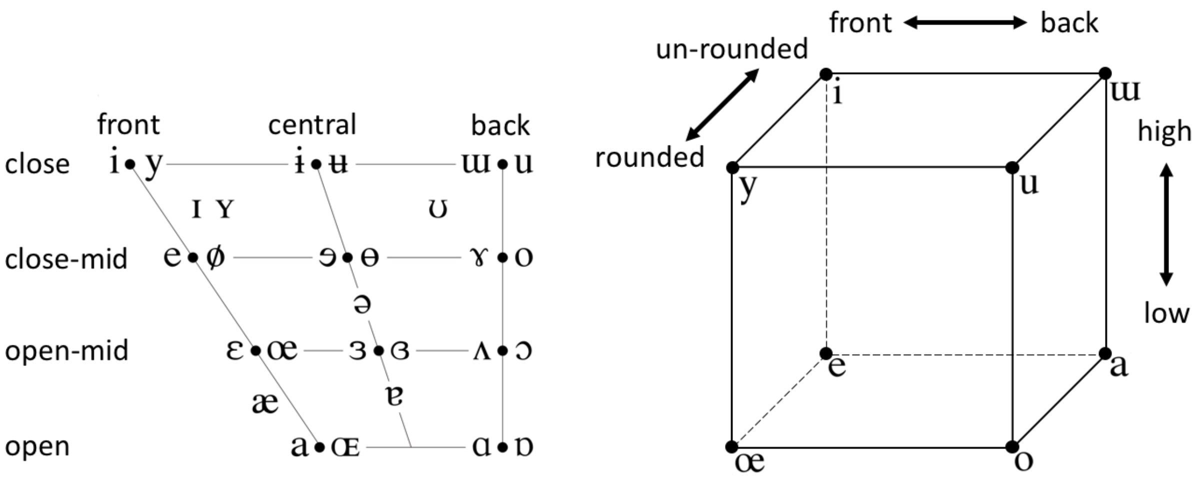


Fig. 3.1: Traditional vowel trapezoid vs. Turkish vowel cube. (Left) In representing the vowels, it is traditional to use a 2D plane, that is based on the shape of the oral cavity, where the two dimensions are the the relative height of the tongue and its advancement (relative frontness or backness) during vowel production. A third dimension, that is the amount of rounding of the lips, is smashed onto the trapezoid, and hence where vowels appear in pairs, the one to right denotes a rounded vowel and the one to left an unrounded vowel (such as the pair /i/•/y/ in which /y/ on the right is rounded). (Right) The Turkish vowel repertoire is comprised of 8 categories which can be realized as a combination of 3 different binary properties that includes rounding in addition to the usual two dimensions: openness and fronting. Therefore, it can be argued that, in Turkish, the rounding feature turns the traditional trapezoid into a cube.

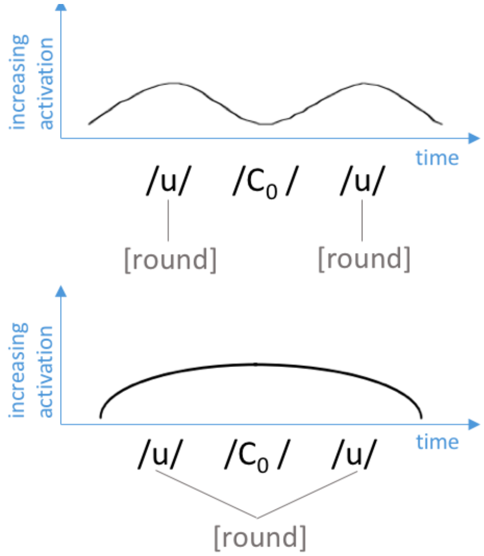


Fig. 3.2: EMG traces of speakers of Turkish vs speakers of 5 other languages. (Top) When speakers of English, in which there is no rounding harmony, articulate words that have the sequence of /u/-consonant-/u/, the electrical activity generated by the muscles around the lips exhibits a 'valley'-'plateau'-'valley' pattern (valleys appear as hills when plotted as activation values). (Bottom) That can be contrasted to the pattern of Turkish speakers when they articulate the words of the same form, during which the lip protrusion that starts with the first rounded vowel remains till the end of the second rounded vowel. Adapted from [Boyce, 1990] and [Kaun, 2004].

same subclass of the last vowel of the word it is added to. In addition to the front-back harmony, a second phenomenon that co-occurs with the former is called rounding harmony. According to this, if a syllable contains an unrounded vowel, the next syllable should contain an unrounded vowel as well, but if it contains a rounded vowel, the next syllable should contain either an open unrounded vowel (/a/ or /e/), or a closed rounded vowel (/u/ or /y/).

Do such rules influence the mental representation of the sounds, and if yes how? It has been shown that two different patterns emerge when comparing the electromyographic traces of native speakers of Turkish with those of English when pronouncing [rounded vowel]-[consonant]-[rounded vowel] sequences [Boyce, 1990]. As native English speakers articulate a word of the above-mentioned form, the electrical activity generated by the muscles around the lip goes up during the first rounded vowel, then down during the consonants and then up again for the second rounded vowel (see the top panel of Fig. 3.2).

As shown in the bottom panel of the same figure, however, Turkish speakers exhibit a different pattern, resembling that of a plateau: it seems like rounding is spreading from vowel to vowel, skipping over the intervening consonant. This experimental finding suggests that whereas English speakers execute two lip rounding movements, Turkish speakers perform only one, suggesting two distinct language-specific phonological representations, reflected by the dissimilar activity patterns in Fig. 3.2, resulting in two dissimilar phonetic behaviors [Kaun, 2004].

3.1.2. Rounding: a third dimension?

Considering the importance of rounding, as a perceptual cue to native speakers of Turkish, is it a fundamentally different dimension inflating the trapezoid into a cube, more than a simple shift in the F1, F2 plane, as it has been traditionally thought?

In order to test this hypothesis, we have designed two different acoustic models of rounding, to be later used to generate synthetic stimuli. In the first model, we implemented rounding as a small and constant difference between the second formants of the rounded and the unrounded vowels, as it is captured in the traditional representation (see the left panel of Fig. 3.1). In the second model, we included an additional temporal gradient to the second formant. In this case, both rounded and unrounded vowels have the same mean second formant, as a contrast to the first model of rounding, but the rounding feature changes dynamically throughout the time course of the vowel, which is no longer treated as a point in the plane, but rather as a trajectory (see Section 3.2 for further details).

On which of the two perceptual cues would Turkish speakers rely to distinguish between unrounded and rounded vowels? If rounding is another dimension, we would expect them to be able to tell apart unrounded and rounded vowels by relying on the temporal changes, even if the vowels have the same mean formant value. Their performance in doing so should be at least as good as their discrimination between the vowels when rounding is projected onto the 2D plane, and represented only by a static difference between the formants.

What about speakers of a language in which rounding is not perceptually salient, like Italian? Would they be able to perceive at all the difference between the sounds when the mean formant value is the same, but the rounding feature evolves through time? Would Turkish speakers perform better at following those dynamic changes that define the rounding property of a vowel compared to speakers of Italian, for whom we do not expect rounding to be a principal component?

3.2. Materials and methods

3.2.1. Stimuli

For each one of the 4 pairs of rounded and unrounded Turkish vowels (see pairs /i/ /y/, /ɯ/ /u/, /e/ /œ/, /a/ /o/ in Fig. 3.3), we created 4 vowel trajectories with the Klatt Speech synthesizer [Klatt, 2013], for a total of 16 stimuli.

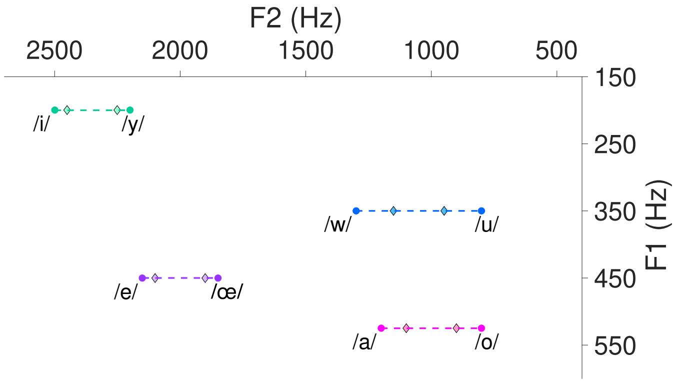


Fig. 3.3: Four rounded and unrounded pairs of vowels of the Turkish. For each vowel pair, we have created two steady and two dynamic trajectories. Diamonds denote F1 and F2 values of steady trajectories. Circles denote F1 and starting (or ending) F2 values of trajectories that are dynamic (see Fig. 3.4).

The 4 trajectories in every pair differed in their F2 values, but have the same F1 values. For each of these 4 pairs, we have created two different couples made up of 2 trajectories. These two different couples corresponding to two different models of rounding: dynamic and steady. In one couple the 2 trajectories each have a steady second formant for 400ms, as seen for the pair /w/-/u/ on the top panel of Fig. 3.4. The F2 and F1 values of the trajectories in the 'static couple' are denoted by the diamonds in between the circles in Fig. 3.3. Indeed, in the steady model of rounding, we define rounding as a constant difference in F2. In the other couple, i.e. the dynamic model, the 2 trajectories had a changing second formant over the time course of the signal; both sounds started with two different but close second formants, but at the end of 400ms, the final F2 value was the initial F2 value of the other one so that their mean F2 was kept the same (see the bottom panel of Fig. 3.4). Therefore for this second couple of trajectories representing the dynamic model of rounding, the rounding feature is expressed in the slope of F2. In Fig. 3.3, the circles show the starting and ending F2 values of the two trajectories in the dynamic couple of every rounded-unrounded pair. We chose the circles somewhat wider apart than the diamonds in order to approximately equalize the discrimination difficulty; although whether we succeeded in doing that and if so for which group of subjects will become clearer in the Section 3.3 where we discuss the results. The F3 values of the 4 vowel pairs were chosen different, in order to make them sound less metallic, but in both conditions it was kept the same and equal for the unrounded and rounded vowels, whereas all the sounds were generated with the same default fundamental frequency, i.e. F0, which represents the pitch.

3.2.2. Participants

We tested 33 non-bilingual native Italian speakers, and 30 non-bilingual native Turkish speakers During the pilot study we conducted, we noticed that the participants who played a musical instrument could spot the differences between the pairs very easily. We therefore eliminated such participants and those that were included in the experiment had no musical instrument playing skills and did not receive any formal training in music.

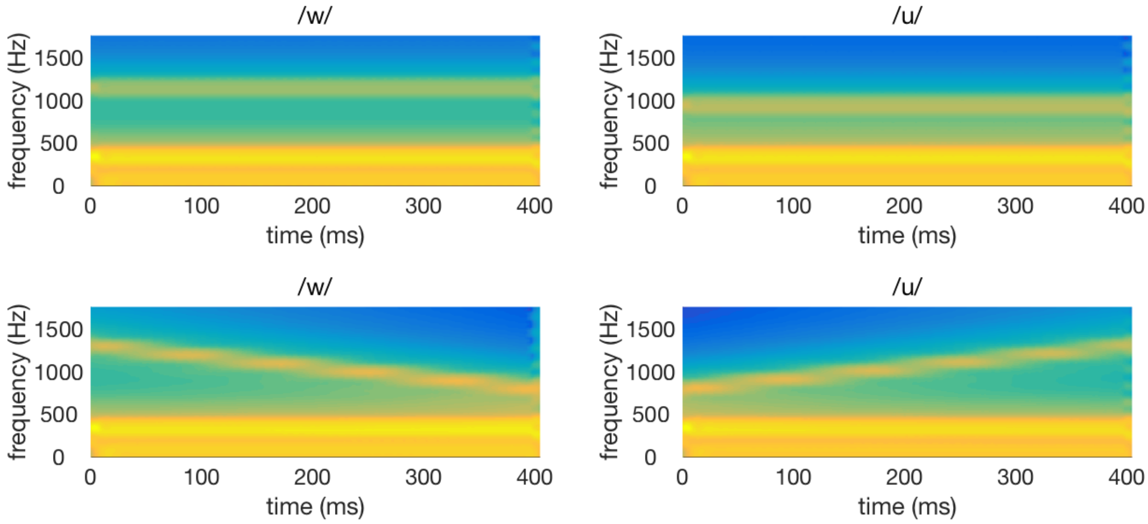


Fig. 3.4: Two different couples of a pair of unrounded and rounded vowels based on their time courses. (Top) Both unrounded and rounded vowels have steady, but different, F2s. (Bottom) Both unrounded and rounded vowels have the same mean F2, but evolving in time downward or upward from one value to another.

3.2.3. Paradigm

There were 160 trials in total. Each trial started with a 1500ms long silence, followed by the presentation of a pair of 400ms long sounds separated by a 200ms silence. In half of the trials, the two sounds were identical; they either had the same steady F2 (5*8 = 40 trials), or the same dynamic F2 (5*8 = 40 trials). For example, a trial with the same sounds in the steady condition could have [/w/ and /w/], both realized as static unrounded vowels, or [/w/ and /w/], both realized as dynamically unrounded vowels —spectograms of which can be seen in the top left and the bottom left panels of Fig. 3.4 respectively. In the other half of the total trials, the two sounds were different; they had either a steady but different F2 (5*2*4 = 40 trials) or a dynamic F2 but with the opposite slope (5*2*4 = 40 trials). An exemplary trial with two different trajectories would be [/w/ and /u/] in the steady condition (the top panel of Fig. 3.4) or [/w/ and /u/] in the dynamic condition (the bottom panel of Fig. 3.4).

The presentation order of the sounds in a pair of two different trajectories (as in [/w/ /u/]) was alternated: so that for half of the trials started with one of the sounds in the pair presented first ([/w/ /u/]) and the other half started with the other sound presented first ([/u/ and /w/]) to account for the directional asymmetries in vowel perception [Polka and Bohn, 2003, Masapollo et al., 2017]. The order of the trials was randomized in such a way that no two consecutive trials had one of the two couples of the same rounded-unrounded pair. The participants were asked to respond if they perceived the two sounds as the same by pressing the space bar. A new trial started in 2 seconds unless they had already responded in case of which a new trial began immediately. During the analysis, late responses up to 100ms after the 2 seconds time limit were accepted as 'same' response. After every 40 trials, participants had a 40 seconds long break.

3.3. Falsification of the dynamic Turkish formant hypothesis

In Fig. 3.5, we can see that participants are able to identify pairs of same sounds accurately, regardless of whether the pairs are static or dynamic. Moreover, as expected, they respond with faster reaction times (see Fig. 3.6 where dashed lines, representing such trials, are lower than the solid lines corresponding to 'same' responses in trials with different trajectories). When they are presented two different sounds, they are more likely to perceive as 'same' the pairs in which two trajectories have the same mean second formant, i.e the trajectories that have dynamic second formants, than the pairs in which two trajectories have different, but very close, mean second formant, i.e the trajectories that have steady second formant. In contradiction with our hypothesis, native Turkish speakers are not better at following the temporal changes in the second formant than native Italian speakers. In fact, contrary to the fact that they use all these 8 vowels, they are more likely to perceive two trajectories that have different mean F2 more as similar than their Italian counterparts.

Do the responses of each mother tongue group depend differentially on different unrounded-rounded pairs? We find that when two same trajectories are presented, both language groups show similar patterns across the 4 pairs independent of the F2 condition. However, we observe some differences in their same responses when two different trajectories are presented, as shown in Fig. 3.5.

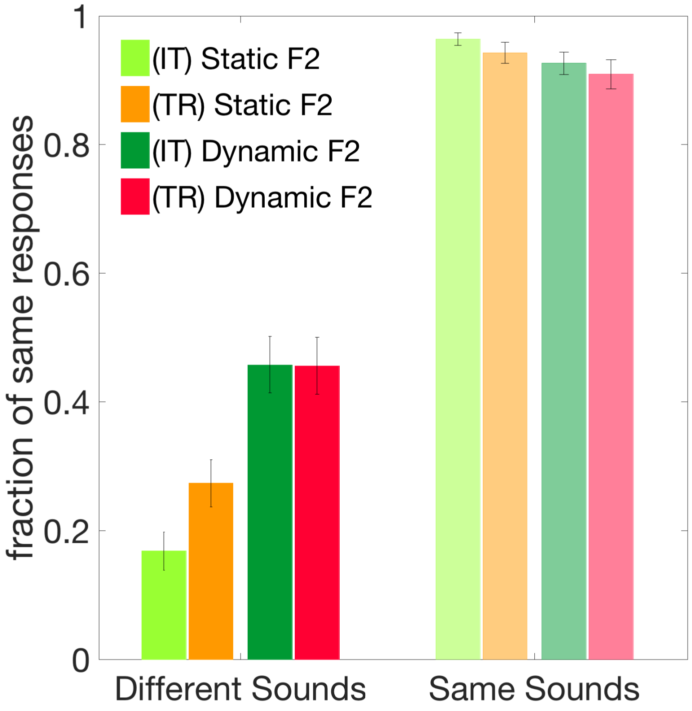


Fig. 3.5: Perception of static and dynamic trajectories by native Italian and Turkish speakers. Pairs of sounds including two same trajectories, either steady or dynamic, were identified as same more than the pairs of two different trajectories. Different trajectories that have the same mean F2 (i.e. dynamic trajectories) were identified more as same than different trajectories that have different mean F2 (i.e. static trajectories) by both mother tongue groups.

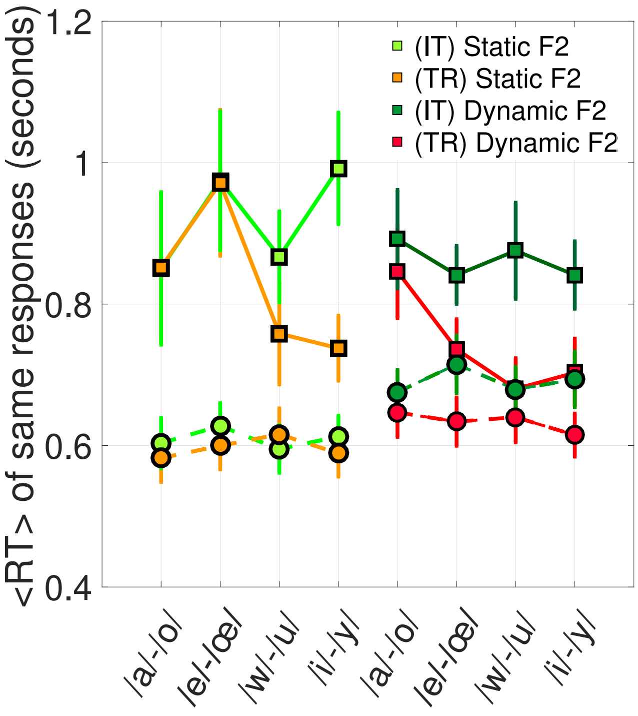


Fig. 3.6: Mean reaction times of same responses across different conditions. Solide lines (and squares) represent mean RTs of trials with two different trajectories, dashed lines (and circles) represent mean RTs of trials with the identical trajectories. It takes more time to respond when two trajectories are different than when they are identical. For the pairs made up of two steady trajectories, native Turkish speakers respond faster to decide /w/-/u/ and /i/-/y/ as identical. The two groups of speakers have the greatest difference in their RTs for the /i/-/y/ trajectories. For dynamic pairs, they make their decisions faster to judge them as same compared to native Italian speakers, especially for the /w/-/u/ trajectories.

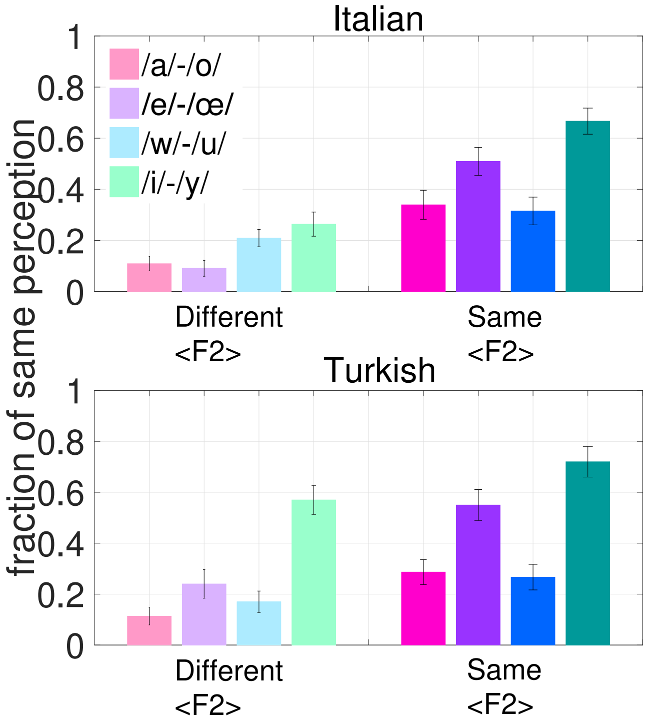


Fig. 3.7: Perception of trials with two different trajectories. Among the static trajectory pairs, the ones that have different mean F2, the major differences between the two mother tongue groups are among the pairs /i/-/y/ and /e/-/œ/. Native Turkish speakers perceive these two pairs, especially the first one, as as more similar than Italians do. On average, Italian speakers have slightly more difficulty in perceiving the distinction between /ɯ/ and /u/ trajectories. Among the dynamic trajectory pairs, the ones that have the same mean F2, both mother tongue groups show similar patterns across the 4 pairs. The dynamic trajectories corresponding to /i/ and /y/, and also to /e/ and /œ/, are perceived as more similar than the other pairs.

Even though the F2 difference between the two steady trajectories in every pair is the same (200Hz, see Fig. 3.3), we see that the relatively high fraction of same responses of Turkish speakers to trials in which two different steady trajectories are presented (as shown in Fig. 3.5) is mostly due to the pair that is in the frontmost position of the vocal tract, i.e. the /i/-/y/ couple (see Fig. 3.7). Compared to the perception of native Italian speakers, Turkish speakers identify these two different trajectories to be closer to each other, and this is accompanied by the shortest reaction times of the judgements they make for this pair (see the average reaction times of same responses for the condition 'Static F2' in Fig. 3.6). The same figure shows that with native Italian speakers, although the /i/-/y/ pair is, on average, perceived as more similar than the other pairs, their reactions are not as fast as the reactions of Turkish speakers.

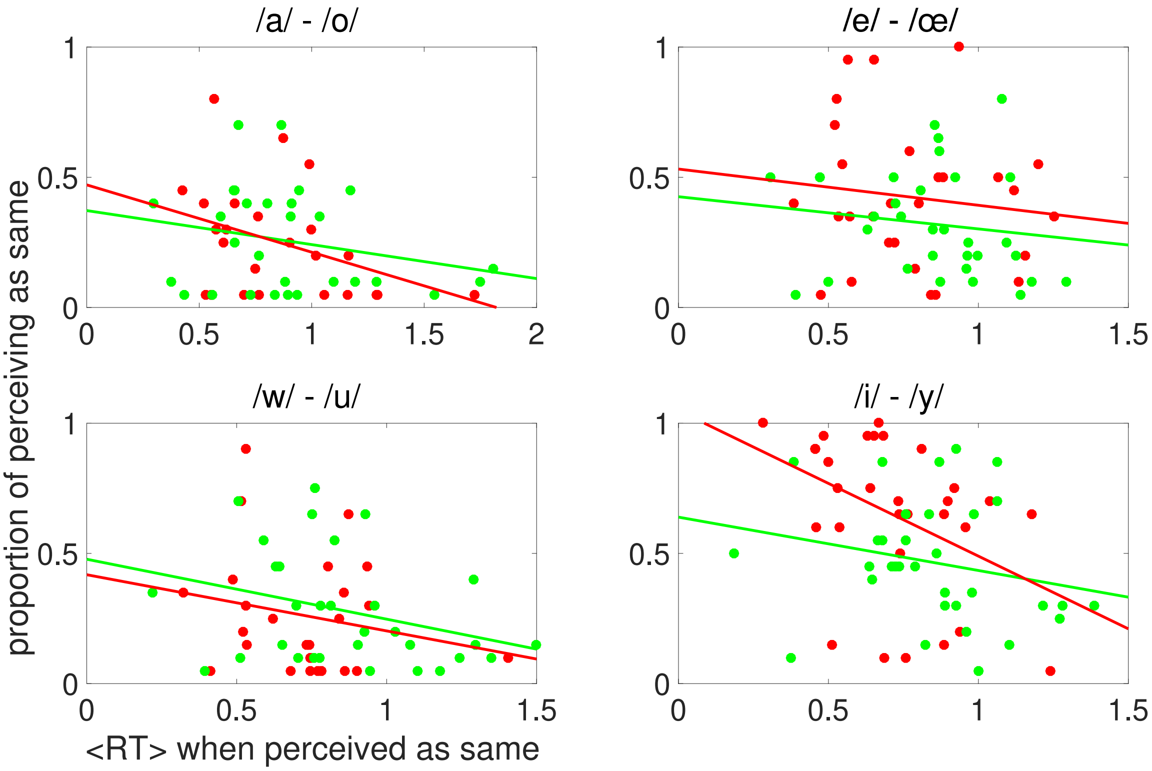


Fig. 3.8: Correlation between the proportion and the reaction time of wrong answers. For different vowel pairs, the scatter of the individual data points (green circles for Italian and red circles for Turkish participants) shows the proportion of 'same' responses (on the y-axis) and the average reaction time for these responses (on the x-axis) when two different sounds were presented. The superimposed lines are the least squares fits. Both Italian and Turkish subjects respond faster when they make errors. The correlation is particularly negative for Turkish subjects when they are presented the pair /i/-/y/.

From the high fraction of same responses and short reaction times, it seems like Turkish speakers are confident in their judgement of perceived similarity among /i/ and /y/. Indeed, Fig. 3.8 shows that, for both mother tongue groups, there is a negative correlation between the average reaction time to respond as 'same' and the proportion of 'same' responses when two different sounds are presented. The slope of the least-squares fit line is the steepest for Turkish speakers when they respond to the /i/-/y/ pair. Not included here is the finding that the same participants spend more time to respond when they make errors after they hear two identical sounds. This observation rules out the possibility that the participants did not execute the task with care. Rather, it seems that the frontal region of the vocal tract is the most resistant and the least affected by the changes in F2.

The /e/ and /œ/ trajectories that have different mean F2 are perceived as slightly more similar by native Turkish speakers, and the least similar by native Italian speakers; however, this difference is not reflected in the reaction times between the two language groups. Turkish speakers spend more time to judge the /e/-/oe/ pair as same than they do for the /i/-/y/ pair despite the same F2 difference. Italian speakers spend slightly more time to identify the pairs /i/-/y/ and /e/-/oe/ as same more than the other two pairs. Overall, interestingly, native speakers of a language that has both vowels of a pair have more difficulty with assigning them into distinct categories than the native speakers of a language that does not have one of the two vowels (in any of the pairs, except the /a/-/o/ pair). Although the two trajectories /a/ and /o/ can be distinguished from each other successfully by both mother tongue groups, it is not the easiest pair to judge, as the reaction times are not the fastest for Turkish speakers, and closer to the reaction times of /w/-/u/ trajectories for native Italian speakers. Turkish speakers spend less time to decide /w/-/u/ to be identical, when they do, compared to Italian speakers for whom they are slightly more similar.

Fig. 3.7 shows, for trials in which two dynamic trajectories that have the same mean F2 are presented, that similar patterns of perception are observed among the speakers of Italian and Turkish. For trajectories, where the rounding feature changes between /i/ and /y/, the fraction of same responses is the highest. Similar to the difference between the reaction times of steady /i/-/y/ trajectories, Turkish speakers judge the dynamic trajectories /i/ and /y/ to be similar faster than Italian speakers (Fig. 3.6). The F2 difference between the two end points of this pair is equal to that of /e/-/œ/ pair (300Hz, see Fig. 3.3), hence similar differences between the reaction times among two mother tongue groups for these two pairs /i/-/y/ and /e/-/œ/, yet the pair /i/-/y/ is perceived as more similar.

When the difference between the two end points is increased up to 500Hz, as it is in the /w/-/u/ pair, the fraction of same responses decreases both for Italian and Turkish speakers. However, the difference between reaction times of the /w/-/u/ pair among two mother tongue groups is the greatest, i.e. Italian speakers spend more time to judge them to be the same, while Turkish speakers give the verdict for these trajectories fastest. On the other hand, despite the similar fraction of same responses, Turkish speakers take more time to decide the pair /a/-/o/ to be the same, although a

400Hz difference between the two end points is not as small as a 300Hz difference in the pairs /i/-/y/ and /e/-/œ/, and not as big as 500Hz difference in the pair /ɯ/-/u/. This increased reaction time is the closest to that of Italian speakers across 4 pairs.

3.4. Discussion

Previous studies have found evidence of an interaction between fronting and rounding resulting in front vowels being perceived as more unrounded whereas back ones as rounded, presumably due to the relative frequencies of front/back ± rounded/unrounded vowels [Lisker, 1989]. Due to the educational training, Turkish speakers are conditioned to make a symmetric classification of their vowels, so that, every front unrounded vowel has also a rounded counterpart. Here we used a novel approach to assess their perception of the rounding dimension in relation to the changes in F2, dependent or independent of time, and compared it with that of native Italian speakers who do not receive an external instruction to make such categorizations.

The results of the rounding experiment suggest that a third dimension, i.e. rounding as a dynamic F2, is not needed and can be surrogated with a static F2 difference. However, our findings do not mean that other 'third' dimensions and/or this dimension in other languages are not relevant or even essential: consider the four tones in Mandarin Chinese for example [Huang and Johnson, 2010].

The 'same' responses for static F2 difference and for opposite F2 slopes, as in Fig. 3.5 and in Fig. 3.7, need not be quantitatively similar, because they have not been strictly calibrated. We chose F2 ranges of all pairs in both conditions based on the series of pilots we conducted with the native Italian speakers, but not with the native Turkish speakers as we only had access to the former group first. However, in the light of our results, it is unlikely that the Turkish mother tongue group would have outperformed the Italian group if F2 had been changing in a narrower range of values. The important result we find is not that the 'same' responses are similar, but that there is no interaction between condition and subject group. This should be tested statistically, although it is obvious by eye.

Another possible source of interference with our results could be the synthetic stimuli we used.

Although Turkish speaking participants confirmed that they heard Turkish vowels in the experiment, which was not told a priori to the participants, perhaps we did not have enough variation of the exemplars for each pair, and it was sufficient for the participants, especially native Italian speakers, to engage in tone processing, but not necessarily vowel processing. To further investigate the rounding feature, it would be interesting to try similar manipulations on human articulated stimuli, which would result in more natural sounding speech.

4

Vowel (or Diphthong) Trajectories

4.1. Movement in vowel space

The study of roundedness perception in the previous chapter, while producing essentially negative results, illustrates the wider options introduced by allowing for movement in vowel space. If a rounded vowel can be distinguished from its unrounded counterpart by the temporal gradient of its second formant frequency, we can formulate the general hypothesis that the temporal variation in any combination of formant frequencies may enable finer distinctions among phonemes than if they were realized strictly as static spectrograms. Given the large variability in the mean frequencies themselves, for example between typical male and female utterances of the same phonemes, additional, dynamical cues that enable discrimination are certain to carry significant perceptual value.

This hypothesis is reinforced by the observation that in many languages it is traditionally assumed that some diphthongs, or pairs of consecutive vowels fused together in pronunciation, carry their own phoneme identity distinct from that of the combination of their constituent values. Acoustically, diphthongs have been characterized as the succession of very brief initial and final segment with constant formant frequencies, joined by a longer intermediate segment in which at least the first and second formant "slide" continuously to move between the values corresponding to the first and second vowel of the diphthong [Gay, 1968, Fox, 1983]. A trajectory, in other words.

In analogy with spatial navigation, therefore, during which activity patterns are thought and sometimes observed to settle into stable states of activity that correspond to a specific location in the environment [Jezek et al., 2011, Lorenzo et al., 2018], we would like to understand how continuously changing vowel dimensions F1 and F2 may be related to discrete sound categories. Is this similar to the case for attractor dynamics? Are these analog-to-digital, i.e. continuous-to-categorical, computations influenced by how different vowels are encoded in the long-term memories of speakers of different languages, which vary significantly in where they 'put their vowels' on such a manifold? How do we navigate the vowel trajectories in this continuous space of vowels? The neural basis of spatial navigation is well studied in rodents and recent advances have even allowed to gain insights into the human brain. Can we use this knowledge to shed light on the mechanisms involved in how our brain operates to navigate in the space of vowels?

4.1.1. Neural representation of spatial navigation

The discoveries of spatial representations by neural populations in Medial Temporal Cortex (MTC) brought a revelation to the neuro-scientific endeavor which till then had largely focused on the neural computations in the very first sensory areas, such as in the seminal study by Hubel and Wiesel who showed the nature of the topographical map in the striate cortex [Hubel and Wiesel, 1959]. It was the discovery of place cells in the hippocampus which clearly demonstrated that it is possible to go beyond sensory-motor systems and start understanding what is going on in the deeper regions of the brain.

Hippocampal neurons that respond preferentially to the specific locations in space are called 'place cells ' (see an examplar place cell with its place field in the Northeast corner of the spatial enclosure in the leftmost firing map in Fig. 4.1) [O'Keefe and Dostrovsky, 1971]. The representation of a spatial environment is achieved through the entire population with different cell ensembles having their own firing locations [Wilson and McNaughton, 1993]. In one synapse upstream of the hippocampus, in the medial part of entorhinal cortex (MEC), another group of cells called grid cells fire not only in one particular field, but each in several fields, reliably, as the animal traverses the entire environment [Fhyn et al., 2008] . The activity of grid cells is presumed to reflect an absolute neural metric for local space, as their positional relationship with each other is maintained in different environments [Moser et al., 2008], irrespective of environment specific contextual details, i.e. *invariant*, in contrast to place cells, whose activity change randomly from one environment to the next —a phenomenon called as *remapping* [Muller et al., 1987]. The defining feature of grid cells is the hexagonal lattice pattern formed by the multiple discrete and periodic firing locations of each neuron that tile the space, as shown in Fig. 4.1 (see the third spike plot from the left on which the vertices of the white triangular pattern at the firing fields is superimposed).

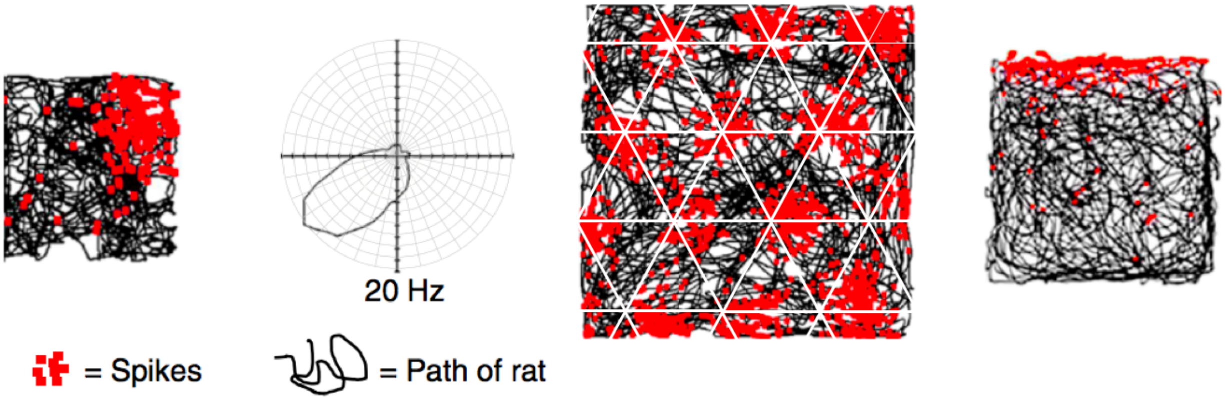


Fig. 4.1: Different neuronal populations that form the cognitive map. Spatial cognition requires having a *cognitive map*, i.e. an internal representation of the outside world [Tolman, 1948]. Single neuron recordings have led to the discovery of various neuronal cell assemblies, each with its own specific function, in the hippocampus and the enthorhinal cortex. The different biological features of each presumably serve different specific functions, and have been found to provide the neural basis of spatial navigation. The activity of a place cell, a head direction cell, a grid cell, and a border cell in environments of different sizes are illustrated in the spike plots from left to right. (Adapted from [Marozzi and Jeffery, 2012])

The grid of each cell is defined by three properties: its spacing (i.e. the distance between the firing fields), its orientation (i.e. the tilt angle relative to a reference axis), and its spatial phase (i.e. displacement in the x and y directions relative to an external point). Along the dorso-ventral MEC axis, grid cells are arranged into a few discrete modules (4-5), in which cells with similar grid scale and orientation but with different grid phase are clustered together [Stensola et al., 2012], as shown for three neighboring grid cells in Fig. 4.2.

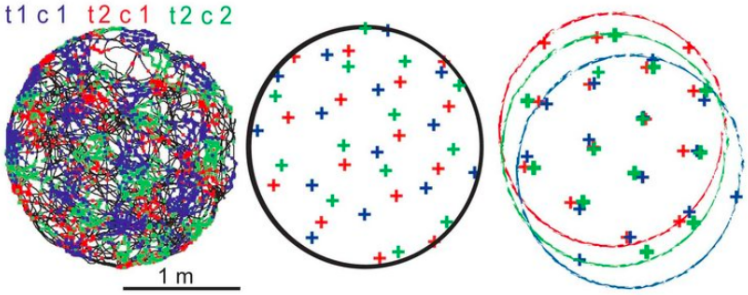


Fig. 4.2: The relationship between co-localized grid cells. (Left) A population of grid cells forms a map of the environment, here the firing fields of only three co-localized grid cells (in blue, red, and green), that are simultaneously recorded from a rat navigating in a closed enclosure (with a radius of 1 meter), are shown. (Middle) The peak locations of each of the three cells seem like they are randomly distributed. (Right) The peaks are offset to visualize similarity in spacing and orientation, but difference in phase. (Original figure taken from [Hafting et al., 2005])

The representation given by grid cell activity is driven mainly by self-motion sensory inputs and is dynamic in nature; when an animal explores alternated corridors, the activity is least correlated at the turning points, but right after the turns, spatial firing fields are significantly similar in all corridors oriented in the same directions [Derdikman et al., 2009]. Moreover, when the animal forages an entirely novel environment, the firing structure, the size of the fields and the distance between fields of grid cells remain unchanged, whereas the phase and orientation are changed so that the new grid is shifted and rotated, but, remarkably, in the same manner within a module [Stensola et al., 2012], [Fhyn et al., 2007]. We discuss the multiplicity of the spatial maps in Chapter 5.

Finally, the neural representation of physical space is not limited to place and grid cells, but enhanced by at least three other separate populations of neurons that coexist with grid cells in the MEC. These neuronal groups are functionally discrete, and hence according to their functional tuning, they are called head-direction cells, border cells, and speed cells (see Fig. 4.1 for representatives of the first two) [Taube et al., 1990, Solstad et al., 2008, Kropff et al., 2015].

4.1.2. Grid-cell like representation in humans

Grid cells were initially discovered in rodents [Fhyn et al., 2008], but were subsequently also found in bats [Yartsev et al., 2011] and, perhaps, in monkeys [Killian et al., 2012]. That cells of a similar nature are found across widely differing species lead to hypothesize that similar cells may also be found in the human brain. However, the invasiveness of procedures to record directly from such cells renders this unfeasible in humans, and more indirect means have to be employed. In particular, Doeller and colleagues discovered a signal suggestive of the presence of grid-cell like representations in humans using functional magnetic resonance imaging (fMRI) [Doeller et al., 2010]. They recorded the BOLD activity of human participants while they explored a virtual environment. The task involved finding and replacing items in their correct locations. The authors hypothesized that if grid-cell firing is present, there should be an effect of running direction with six-fold rotational symmetry, reflecting the differences between running aligned (bigger signal) or misaligned (smaller signal) to the grid orientation, and given that the orientation of the grid cells is similar across large regions of the EC, meaning that the grids tiling the environment are offset but they are not rotated relative to each other, as we previously explained (see Fig. 4.2), this effect should be detectable from the bulk population activity, as a proxy for grid cells (see the top panel of Fig. 4.3).

To test whether fMRI signal is modulated by navigation direction with a 60°periodicity, the data was divided into two, and from the first half, the participant-specific grid orientation in EC was obtained. The modulation of the sinusoidal regressor aligned with it was then derived from the second half of the data. A 60°modulation was most significant in the right EC, but surprisingly and different from what is observed in rodents, this modulation was present in a variety of other additional regions including posterior parietal, lateral temporal, medial prefrontal and medial parietal cortices. They carried out further controls with 45°(8-fold) and 90°(4-fold) symmetry which were not significant. The average fMRI signal relative to the baseline of the voxels in the right EC that displayed 6-fold rotational symmetry is shown in the bottom panel of Fig. 4.3.

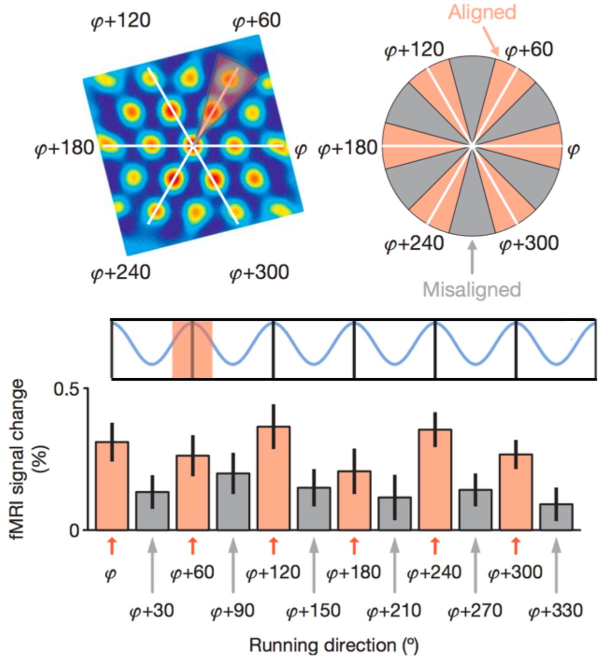


Fig. 4.3: Detecting grid cell activity in humans. (Top) On the left, it is the spatial autocorrelogram of a typical grid cell showing the three main axes of the grid (white lines) and in red a 30 sector aligned with the grid. φ is the grid orientation to which the grid here is aligned. On the right, an illustration of aligned (red) and misaligned (grey) with the grid. (Bottom) The average fMRI signal over the entire time series of all voxels in the entorhinal ROI for all directions of aligned (red) and misaligned (grey) (fast) runs relative to baseline shows sinusoidal modulation of activity by running direction with six-fold rotational symmetry (Adapted from: [Doeller et al., 2010])

Very recently, further supporting evidence has been revealed by two independent studies with different recording techniques. In one study, the participants viewed pictures containing both indoor and outdoor scenes. The authors investigated the basis of grid-like coding during free viewing of natural scenes by combined recording of magnetoencephalography (MEG) and eye-tracking activity from healthy humans, and by simultaneously recording intracranial electroencephalography (iEEG) and eye-tracking data with depth electrodes in the entorhinal cortex of one epilepsy patient [Staudigl et al., 2018]. In the second study, they recorded only iEEG activity from neurosurgical patients as they performed a virtual navigation task [Maidenbaum et al., 2018]. Analyzing different frequency bands, both studies showed hexadirectional modulation of oscillatory power.

4.1.3. Grid cells for conceptual spaces? [1]

The experiments conducted by Doeller et al are suggestive of the presence of grid-like cells in humans. Further studies are needed to ascertain that apart from their hexagonal symmetry, such

[1] For the review article with the same title, see [Kriegeskorte and Storrs, 2016]

grid-like cells possess similar properties to those of grid cells. In humans, an important question that arises is whether their functional contribution is restricted only to the representation of physical spaces or whether such neural hardware can represent more abstract spaces.

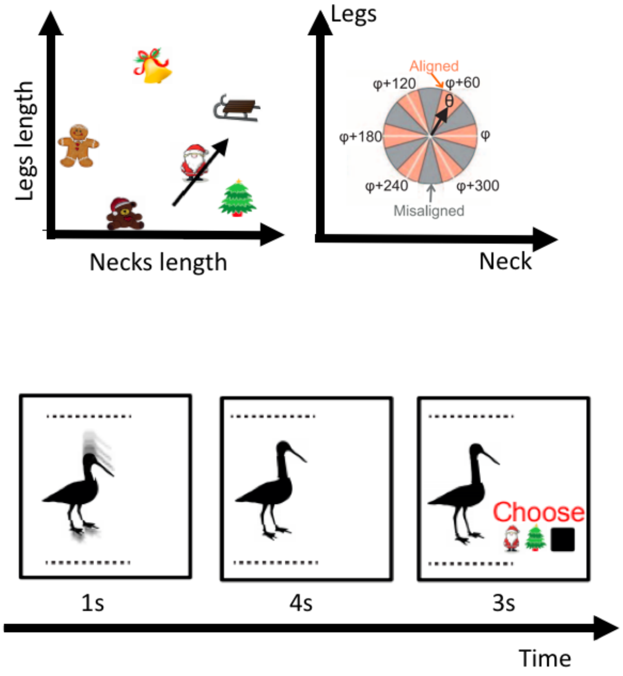


Fig. 4.4: Experimental paradigm of Constaninescu et al. (2006). (Top) Morph trajectories in a 2D plane were defined by the ratio of leg length change over neck length change of bird shapes. Some points in the space were randomly associated with Christmas symbols which participants had to learn by navigating the space. Similar to the experiment of [Doeller et al., 2010], the authors hypothesized that a grid-like representation would entail modulation of the fMRI activity when the trajectories are aligned to the grid orientation (see the trajectory with the direction angle θ). (Bottom) Participants were tested on their navigation skills, for which they had to watch an initial bird morph for some seconds after which they were asked to imagine the rest of the morphing trajectory as well as choose the symbol associated to the final point of the morphing trajectory if there were any. (Adapted from: [Constaninescu et al., 2016])

One experiment which has explored this possibility is the study by Constaninescu et al. (2006). In their study, they constructed such a ‘ conceptual space ’, or a *bird space*, that could be organized into a mental map, arguably similar to the way the concepts are represented in the brain [Constaninescu et al., 2016]. Participants were trained to navigate this space by showing them trajectories in which a bird was continuously morphed from one shape to another. The two dimensions were the length of legs and length of neck of a bird, and the ratio of the former over the latter defined the direction of morphing. The morph trajectories would sometimes pass through locations in the space which were arbitrarily associated with Christmas symbols, and the participants had to learn the association between the bird shapes and the symbols (see the top panel of Fig. 4.4). During the fMRI scanning, on each trial, participants watched an initial bird morph in a random direction for a second, then they were instructed to imagine the morphing to continue in that same direction for further four seconds. To complete the trial, participants had to choose the Christmas symbol that would have been encountered along the morph trajectory (see the bottom panel of Fig. 4.4).

Similarly to what had previously been observed in the grid cell system of rodents, the fMRI signal recorded during this task was shown to be modulated by the morphing direction with a 6-fold periodicity. Moreover, this 6-fold periodic signal was found in a network of regions that overlapped with regions activated during spatial navigation. It seems then that the grid cell system may be able to encode, in addition to physical space, more abstract spaces, even in an artificial task of the type designed by Constantinescu et al.. Can it be that such a code, beyond representing physical space, is a way of representing two-dimensional abstract spaces in general?

4.1.4. How do navigate vowel (diphthong) trajectories?

Can this grid-like code extend to vowels, as we navigate their naturally continuous and abstract quasi-two-dimensional space every few tens of milliseconds, speaking and listening to others? Comparing the position of standard vowels across languages, in Fig. 4.5, we notice that while in Italian they are sitting closer to the edges, leaving the central region empty, British English has richer sets of monophthongs, which ‘tile’ also the inner space. This crowding of familiar acoustic objects likely distorts local geometry, more than in a language with an empty center. But for languages like Italian there is nothing much, in the middle.

The dispersion-focalization theory (DFT) for vowel systems postulates that vowels tend to place themselves at approximately equal distances from each other on the F1,F2 plane with permissible articulation boundaries [Schwartz et al., 1997]. This mean a phonological inventory settles into an optimal arrangement by minimizing an energy function that is a weighted sum of two terms: first one is the *dispersion*, that is maximization of the auditory distances between vowels, and the second one is the *local focalization*, that is maximization of the importance or perceptual salience of focal vowels[2]. This solution of vowels arranging their positions in the space is similar to the self-organization of multiple grid fields into a triangular grid pattern with no external drive, but solely due to the firing rate adaptation that is similar to the dispersion term [Kropff and Treves, 2008]. In a large space with a large number n of vowels they would be close to the vertices of a triangular grid, but since n is only 3, 5, 7,11, max 15 or so in a constricted space, their position depends heavily on the number. If there is a schwa-like vowel, as there is in English (see Fig. 4.5) to acquire similar distances from the vowels pushed at the borders it would tend to extend in a third dimension, effectively deforming flat 2D space into something with curvature. Should we expect, then, the central region to be conceived as a flat arena, in languages like Italian without standard vowels in the center? Would then neural activity demonstrate hexagonal modulation? We

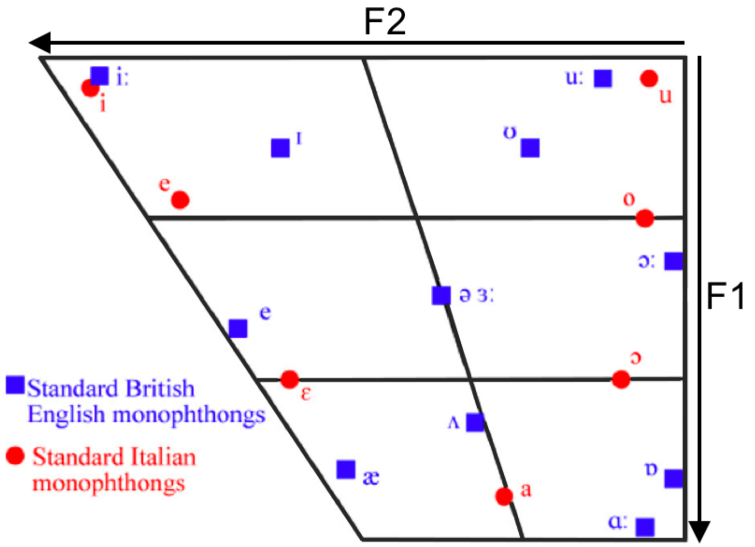


Fig. 4.5: Vowel chart showing the monophthongs of Standard Italian and Standard British English. British and Italian vowel spaces with different number of vowels in different positions. Can we find grid-like cell activity during the perception of sounds in the center which is untouched in Italian and has most likely minimum local curvature? (Adapted from [Bassetti and Atkinson, 2015])

[2] This functional explanation aims to answer why human vowel systems are the way they are, but even for a fixed number of vowels, it does not tell why would one language pick one solution over another, and another language a different one. How the vowel inventories have become this way is address by [de Boer, 1999] who shows vowel systems emerge as a consequence of self-organization through computer simulations.

know a lot about the neural representations of the flat empty arenas used in many rodent navigation experiment, and linking the two would allow us to begin to understand the neural operations our brain carries out on phonemes.

Through an extensive training procedure, in order to get Italian subjects to iron out their perception to be flat in Bark coordinates (see Section 4.2.1 for an explanation of Bark scale), we aimed to search for hexagonal modulation as a function of the direction of change of the first two formants, over two hundred milliseconds, with EEG, which has the necessary temporal resolution.

4.2. Materials and methods

4.2.1. Stimuli for the EEG experiment

In order to reveal a potential 6-fold modulation across directions in vowel trajectories, and compare it, e.g. with a 5-fold one, which we have included as a control, we have created, using the Klatt speech synthesizer [Klatt, 2013], 30 vowel/diphthong (see Chapter 1.2 for the definition of diphthong) trajectories in the 'empty' central circle centered at (10, 4.5) Barks, with radius of 1 (see Fig. 4.6).

The Bark scale is an auditory scale where equal steps in formants correspond to approximately equal steps of perceptual distance; it is roughly linear below 1000Hz and becomes more logarithmic above. Conversion from Hz to Bark is given by the formula [Zwicker, 1961] [3]:

$$Bark(f) = 13arctan(0.00076f) + 3.5arctan\left(\left(\frac{f}{7500}\right)^2\right) \tag{4.1}$$

where f is the frequency in Hz. Using this formula, we first calculated Hz values for Bark values between [0.5,25] in steps of 0.5, and then used linear interpolation to refine the approximation because Klatt works with Hertz scale, and the formula does not permit a simple inverse form.

[3] Based on psychoacoustic experiments, the front end of the human auditory system, the cochlea, can be thought to consist of 24 bank of filters each have a width of one Bark.

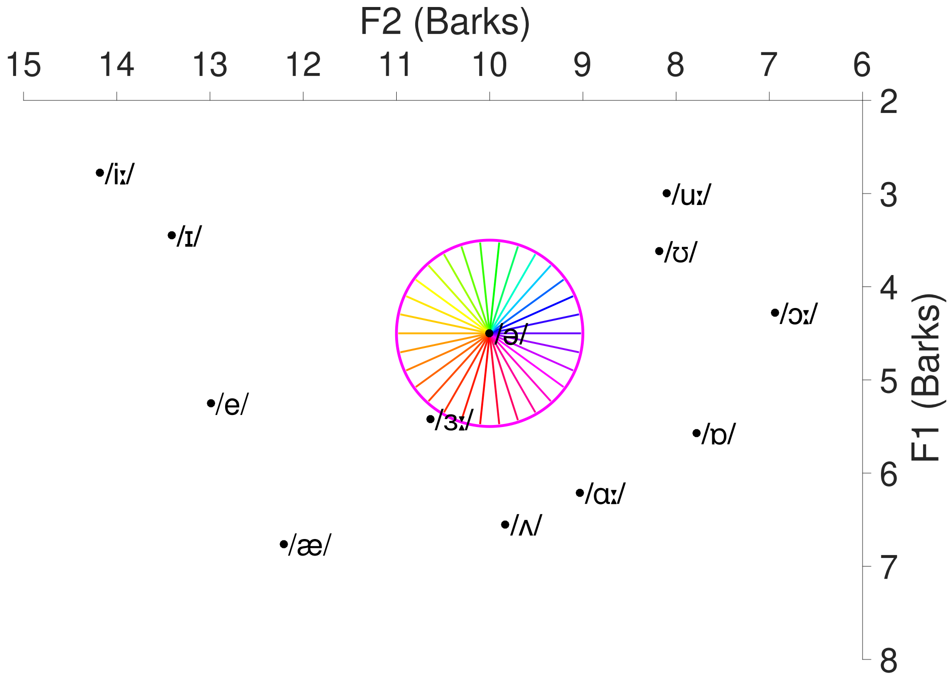


Fig. 4.6: Phoneme trajectories created on the central region in a Bark scale. In this vowel space, where RP British English monophthongs are laid, all 30 trajectories we have synthesized pass through the mid-central vowel that does not exist in Italian. The positions of the vowels are taken from the data by [Wells, 1982] and [Roach, 2004] and converted to Bark scale where equal steps are thought to correspond to approximately equal steps of perceptual distance.

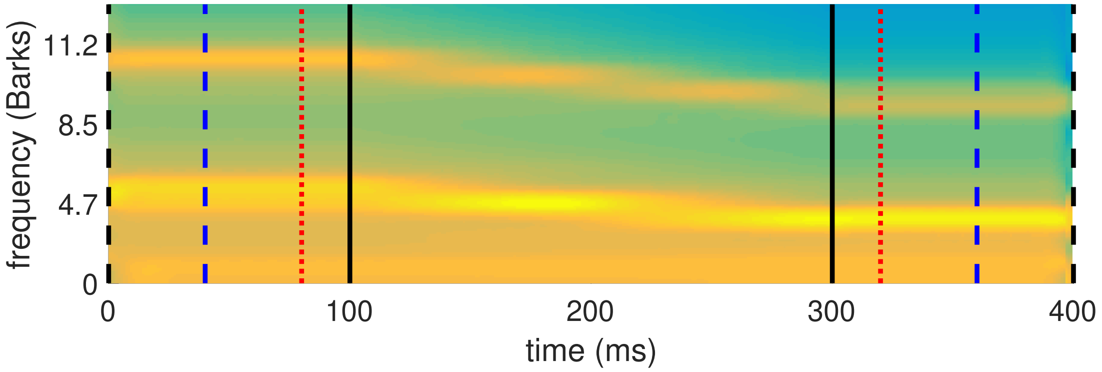


Fig. 4.7: One example of a sound trajectory on the vowel wheel with three different lengths of two steady parts and one same dynamic part. The signal between the black, blue, and red dashed lines depict a trajectory that has two 100ms, 60ms, and 20ms long stable parts respectively. All three variants of each trajectory have a 200ms long dynamic part visible between two solid black lines. During training, two versions with longer (100ms and 60ms) steady parts were presented. Testing was on the discrimination of the shortest version of each trajectory.

Each of these trajectories has steady start and end parts, meaning that during these portions of the signal F1 and F2 values are constant over time, and one dynamic part in the middle during which the signal evolves from the starting phase to the end phase over 200 milliseconds. One such trajectory is shown in Fig. 4.7.

4.2.2. The EEG experiment (The wheel of vowels)

22 native non-bilingual Italian speakers with no or very limited second language skills participated in the EEG experiment. The experiment consisted of 3 blocks of training and testing session pairs. After every training session during which the participants had to learn how to navigate in the central region, they were tested on their newly acquired skill in the test session. Participants had a 40 seconds long break in between every two blocks.

Throughout the three training sessions, the participants experienced 10*30 (300) trajectories in total; half of which were *long trajectories* that have 2*100ms long steady parts (i.e., 400ms in total length), and half of which were *intermediate trajectories* that have 2*60ms long steady parts (i.e., 320ms in total length) as illustrated with black and blue dashed lines in Fig. 4.7.

The 10 presentations of 30 trajectories were distributed unevenly across the three sessions of training in a way that the first training provided the most extended exposure to the participants to facilitate their initial learning. There were 5*30 (150) trajectories in the first training session; where 3*30 (90) were of the long and 2*30 (60) were of the intermediate ones. Every other training session was shorter than the previous one to help the participants to sustain their attention. The second training session had 3*30 (90) trajectories; 1*30 (30) of which were long, and 2*30 (60) of which were intermediate ones. The last training was the shortest session in which 2*30 (60) trajectories were presented; 1*30 (30) of them were long, and the other 1*30 (30) were intermediate.

Along the 2D vowel space, the participants were guided during training by being shown a color square in a hue extracted from a continuous color wheel (see Fig. 4.6), which was randomly rotated for each participant, so that phoneme trajectory-color associations could be established, and they would be different across participants. Every training trial started with 1.5 seconds of silence, followed by the presentation of the sound trajectory. 200ms after the end of the sound, the associated color square appeared and stayed on the screen further for 1 second.

In order to have them engaged in a mental exploration of the vowel space, they were tested on their discrimination of *short trajectories* (i.e., 240ms in total length as shown between the red dashed lines in Fig. 4.7) for which they were expected to extrapolate across the steady start and end parts, which they heard for only 20ms each, similar to the task the participants did in navigating the bird space in which they had to imagine the morphing to be continued in that same direction of an initial bird morph they watched [Constaninescu et al., 2016]. Every testing trial started with 1.5 seconds of silence, followed by the presentation of the short sound trajectory. 200ms after the end of the sound, the response screen appeared with three color options one of which was always the correct answer. The proximity of the choices increased over three different testing sessions (first at 0° +/- 120°, then at 0° +/- 72°, finally at 0° +/- 48° of each other, with the correct choice equally likely to be in each of the relative positions, left, central or right. This way the participants could refine their associations and become increasingly precise in assessing trajectory direction. All response choices associated with one of the sounds in different sessions are illustrated in Fig. 4.8. One test session had 3*30 (90) trials, and the performance on the wheel was assessed over 3*90 (270) trials.

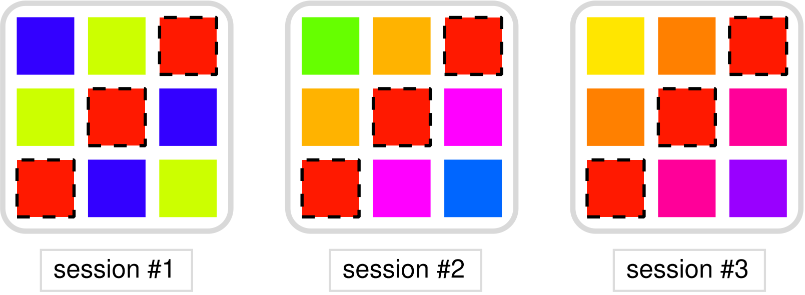


Fig. 4.8: An illustration of response options with increased proximity through different sessions. The correct response is framed with dashed lines, and it is equally likely to be in each of the relative positions, left, central or right. The first testing session is the easiest because the options are quite apart from each other (0° +/- 120°). As the participants receive more training, the proximity of the choices increased so that the participants could be challenged to make finer distinctions during their navigation along the wheel. The options are at 0° +/- 72° in the second, and at 0° +/- 48° of each other in the third testing session.

The participants could evaluate their learning progress with the quiz questions they were asked during the training. After every 30 trials, the training was interrupted with 3 questions that resembled the test trials with a significant difference that not short, but long (400ms) sounds were presented. There was 1 quiz question per trajectory, i.e., 30 questions in total. The color options of the quiz trials were as close to each other as the ones given in the following test session so that the participants were prepared to experience the same level of difficulty during the test. When the participants made a choice, they were given feedback as correct or wrong, but they were not told the right option in the case of a mistake they made.

4.2.3. The shade discrimination experiment

Before the EEG experiment, all the participants were required to participate in a psychophysical experiment in which they performed a shade discrimination task which helped them to acquire awareness of the fine differences of the neighboring hues that were associated with the vowel trajectories.

At each trial, the participants first saw a fixation cross on the screen for 2 seconds, then a pair of squares with the same color or two different colors separated by 12° were presented. The participants had 2 seconds to respond if they perceived the hues of the two squares as identical. A new trial started after the 2 seconds limit unless a choice had already been made. There were 30*2 (60) trials in one session and 60*3 (180) trials in total. Half of the trials included the pairs with the identical color, and the other half included pairs with the neighboring colors. Participants had a 40 seconds long break in between every two sessions.

4.2.4. EEG data collection and preprocessing

EEG data were collected in a sound-proof booth. The stimuli were presented at a comfortable and constant volume from headphones. The brain activity was recorded with a 64 channel BioSemi ActiveTwo system (BioSemi Inc., Amsterdam, Netherlands) at a sampling rate of 1024Hz. A Common Mode Sense (CMS) active electrode was used as the reference, and a Driven Right Leg (DRL) passive electrode was used as the ground. Two external electrodes placed on the right and left of the outer canthi, and one external electrode placed under one eye were used to obtain horizontal and vertical electrooculograms (EOG). Two additional electrodes were placed on the left and right mastoids. Individual electrode offsets were kept between ±30 μV. Participants were requested to minimize movement throughout the experiment except when they had a break.

EEG data preprocessing was performed with EEGLAB toolbox [Delorme and Makeig, 2004]. Offline data was imported by reference to the average of the mastoids as common reference averaging is not preferred for studies of auditory evoked potentials [Khalighinejad et al., 2017], and then band-pass filtered (0.1-30Hz). Following the segmentation of the EEG data into 625ms long epochs starting at around 200ms before stimulus onset and 185ms after stimulus offset, bad channels were discarded using the EEGLAB pop_rejchan function [Delorme and Makeig, 2004]. Trials containing extreme values (±200 μV) were eliminated. On average 7% of data was removed for each subject. Independent Component Analysis (ICA) was used to remove eye blinks and muscle artifacts [Delorme and Makeig, 2004, Chaumon et al., 2015]. At this point, trials with correct and wrong answers were separated.

There were on average 5.36 trials across subjects per trajectory in the final data set of the correct responses. For the clustering (in Section 4.4.1) and ERP analysis (in Section 4.5), the data was divided into the desired conditions, and then it was pruned by randomly discarding trials to ensure the same amount of trials per condition. Finally, missing channels, fewer than 10% of all channels, were interpolated, which was followed by a baseline correction with a reference interval of 200ms before stimulus onset. The resulting dataset of each condition had the data of each participant averaged over trials.

4.3. Behavioral results

4.3.1. Reasonable performance all around the wheel

As Fig. 4.9 shows, the results of the shade discrimination experiment ensures that the participants had no major issues in discriminating among 30 colors, which otherwise could have created a problem for their sound trajectory discrimination. It is true that they are more challenged with perceiving differences between two shades of red and two shades of green, however these were 12° apart (whereas for sound trajectory classification they were at least 48° apart) and also, since the color wheel was rotated across participants, colors were linked to random trajectories.

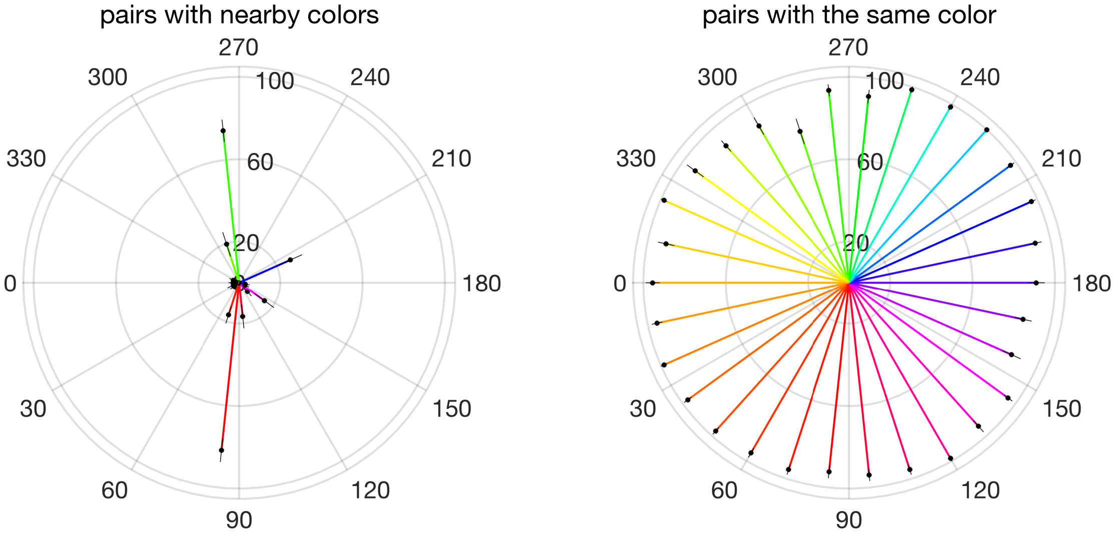


Fig. 4.9: Successful discrimination of nearby colors. The participants are able to notice subtle differences between the hues as reflected in their mean percent 'same' responses for identifying the pairs with the same color (right polar plot), and the pairs with the nearby color (left polar plot) as same. Two shades of red and two shades of green are difficult to discriminate. Error bars denote the SEM across subjects.

The performance was roughly constant across 9 presentations of 30 trajectories in 3 testing sessions (see Fig. 4.10) as desired. Although the task became more and more difficult at each session, gradual training helped the participants to maintain the level of navigation they could achieve. Gradual learning is reflected in the slight increase in the second session in which the second three presentations (presentation # 4,5, and 6) occurred. On the other hand, a slight decrease in the last three presentations (presentation # 7,8, and 9) could likely be explained by the difficulty of the last session due to the increased proximity of the choice options, and some mental fatigue by the participants.

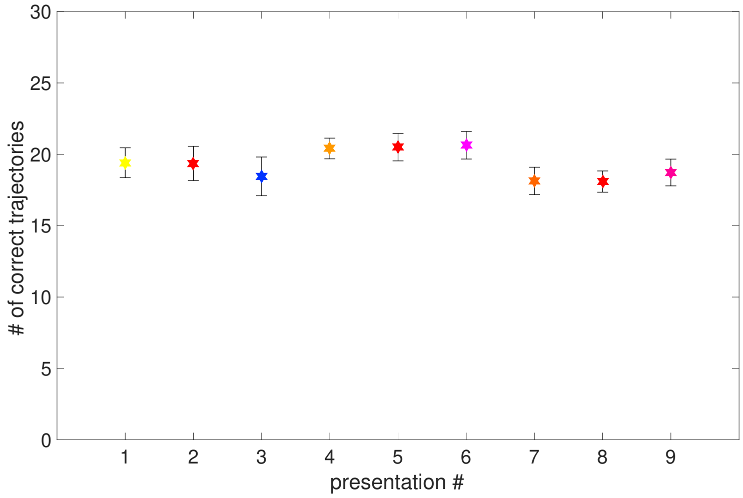


Fig. 4.10: Learning throughout 9 presentations of 30 trajectories. The number of trajectories successfully identified is almost constant over time, with a slight increase in the second session (presentation # 4,5, and 6) possibly due to learning, and a slight decrease in the last session (presentation # 7,8, and 9) perhaps due to fatigue and increased difficulty.

Equally good discrimination of all 30 trajectories was aimed for, in order to reveal a possible grid-like representation of trajectories along the vowel wheel in the middle of the vowel space. The top plot in Fig. 4.11 shows that this is approximately achieved, as reflected in the quasi-circular proportion of correct responses. The inner circle has the radius of the average correct response across 30 trajectories. The responses diverge from the circle especially in the region between 240° and 300° degrees. This region is where the initial 20ms of the auditory stimulus has the formants in between the vowels $/i/$ and $/u/$, where there are no other standard vowel categories in Italian, whereas the final 20ms would correspond, if extrapolated outside the wheel, to the Italian standard vowel /a/. Better performance in those trajectories is accompanied by shorter reaction times, as shown in the bottom plot in Fig. 4.11, in which the inner green circle denotes the average reaction time of correct behavior. Interestingly, the better performance is not entirely symmetric, i.e., the trajectories, which end in that region and start from the opposite bottom part between 60° and 120°

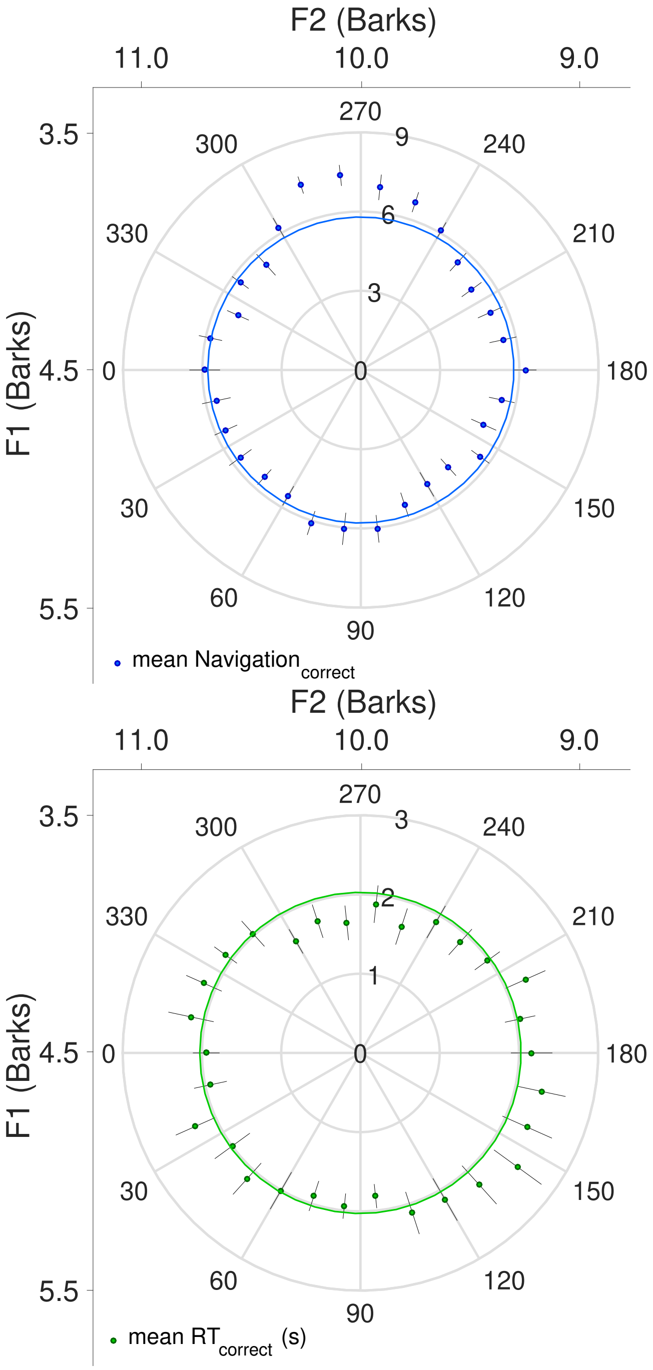


Fig. 4.11: Quasi-circular average performance on the vowel wheel. The top polar plot shows the mean correct response of a group of native Italian speakers with no or little second language experience. The inner circle in blue has the radius of the average of 30 trajectories, slightly less than 6 correct responses out of 9 per trajectory. With training, a somewhat uniform behavior is achieved, but with slightly better performance between 240° and 300° degrees, that is, in the region of the trajectories starting from the empty space between $/i/$ and $/u/$. Reactions to those sounds are slighlty faster as seen in the bottom polar plot in which the inner circle in green shows the average reaction time, in seconds, across all trajectories. The bars denote the SEM.

degrees, are not that well discriminated. If we look at the correct responses session by session, we observe no dramatic difference in the behavior as it can be seen in Fig. 4.12. Trajectories between 270° and 300° degrees are identified better and in a shorter time in all sessions, but the increased performance between 240° and 270° is mostly specific to the first session, where the two wrong choices are easier to discard. It is also observed that it takes more time to recognize the trajectories starting from the region between 150° and 210° degrees, and the correct responses are particularly low, there, in the first session. That portion of the wheel is close to space where vowels /o/ and /ɔ/ are.

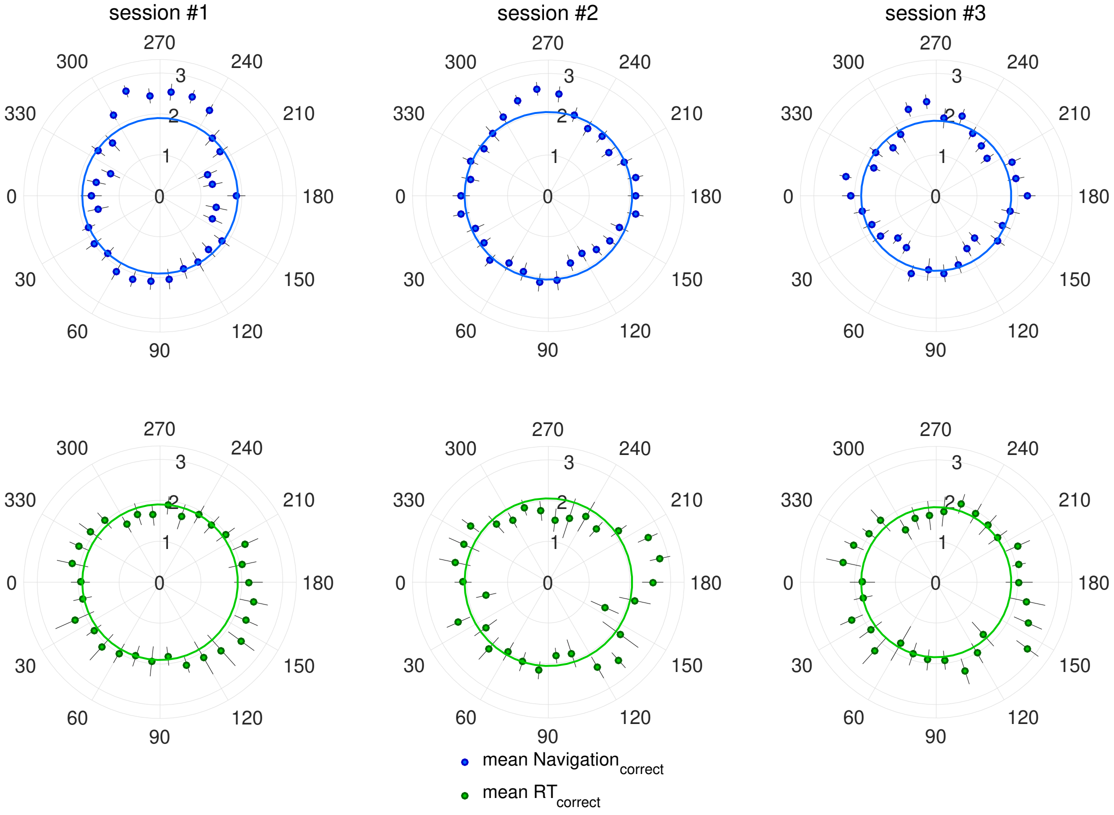


Fig. 4.12: Performance across different testing sessions. The same measures as in Fig. 4.11, but here shown session by session. There were no substantial changes in the perception; its quasi-circular shape is preserved both in the correct responses (top plots) and reaction times of these (bottom plots). The performance for the trajectories between 240 ° and 300 ° was better and faster, but progressively less so with further training. The bars denote the SEM.

4.3.2. Perceptual periodicity

We have first checked for any periodicity displayed in the behavior, to make sure a possible hexagonal or any other symmetry in neural activity is not due to the way the perceptual wheel is deformed. To do this, we can approximate the correct response of each participant on the phoneme wheel, that is $f(\theta)$, by a Fourier expansion:

$$f(\theta) = \sum_{n=1}^{15} (a_n cos(n\theta) + b_n sin(n\theta)) \tag{4.2}$$

where θ is the trajectory angle. Converting this into the cosine form allows the magnitude and phase of each Fourier component to be readily visible:

$$f(\theta) = \sum_{n=1}^{15} A_n cos(n\theta - n\varphi_n) \tag{4.3}$$

where φ_n is the appropriate phase, or participant-specific orientation, that maximizes A_n:

$$\varphi_n = \frac{1}{n} tan^{-1}\left(\frac{b_n}{a_n}\right) \tag{4.4}$$

The better one of those cosines (with its specific phase) approximates the response, the more the responses can be said to show the periodicity of that cosine. Fig. 4.13 shows the weight A_n of each cosine, which we call ‘fold’, up to 10, where the first four components have higher weight, but with none of them significantly greater than the others.

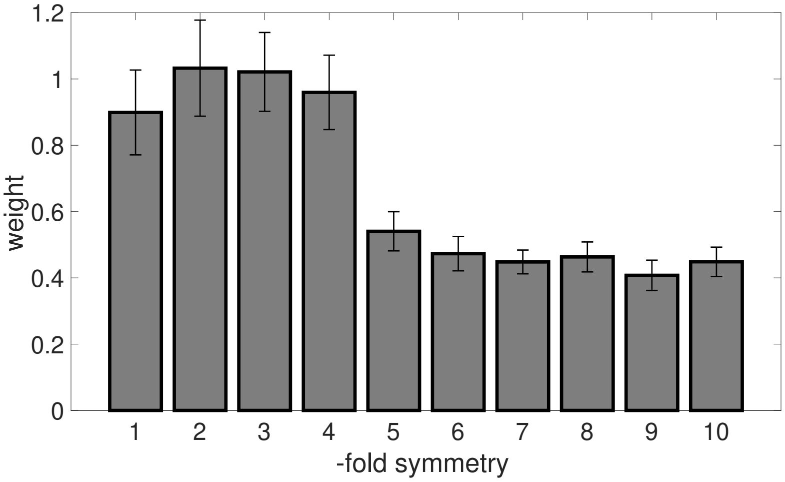


Fig. 4.13: No specific perceptual periodicity in relation to phonemes. Prior to EEG analysis, we checked for the periodicity in behavior. Perception displays no particular symmetry. The bars denote the SEM across participants.

We carried out a similar control test by fixing the color location, as shown in Fig. 4.14, where now $f(\theta)$ is the population mean correct response at the color-based trajectory with angle θ. Here each Fourier coefficient is calculated from that $f(\theta)$ with a population phase φ_{pop} that maximizes it, because rotating the responses for each subject fixing the color location would yield the same maximum weight as in Fig. 4.13, but with different participant-specific phases. In relation to color, we found no specific symmetry that better approximates the responses.

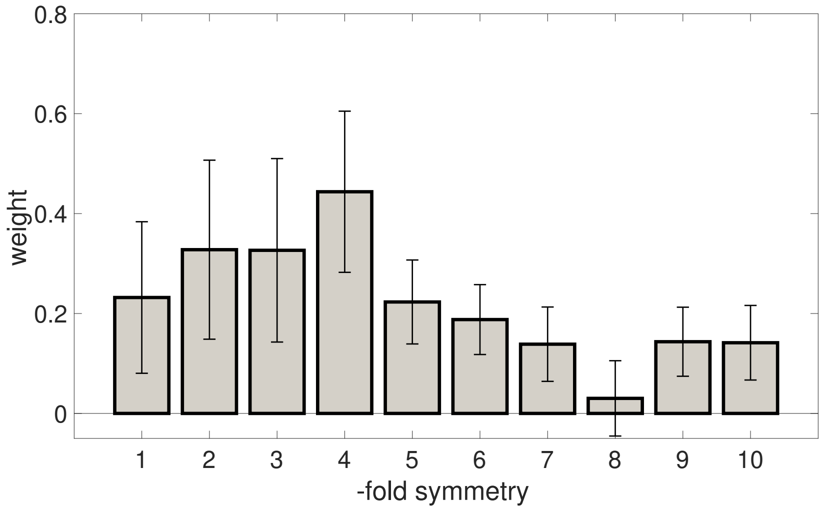


Fig. 4.14: No specific perceptual periodicity in relation to colors. The same control as in Fig. 4.13, but here weights are calculated from the responses that are rotated to match the color locations. The performance in relation to colors does not show a preferential symmetry. The bars denote the SEM calculated across participants with the population phase φ_{pop} that maximizes the weight calculated from the population mean correct behavior.

4.4. Looking for hexagonal modulation of EEG signal

4.4.1. An exploration of periodic processing across the wheel

We employed nonparametric clustering analysis as a method of systematic exploration of the neural correlates of navigation along the wheel, as there are many time points, electrodes, and possible combinations of trajectories.

4.4.1.1. Clustering analysis

This method was first introduced by [Bullmore et al., 1999] and implemented in the FieldTrip toolbox for EEG and MEG analysis [Oostenveld et al., 2011] as a solution to the 'multiple comparisons' problem. As the EEG signal has a spatiotemporal structure, any real effect should not be isolated at a single electrode and time point, but it should be observed over many electrodes across adjacent time points, so that we can exploit the spatio-temporal continuity. Instead of looking for differences between two conditions in a point by point manner, which in the end will lead to a great number of comparisons, this method uses the trick of grouping together adjacent spatiotemporal points and testing whether that group is observed by chance based on the cluster statistic.

For every point in time and space, the EEG signal of 2 conditions is statistically compared by means of dependent samples T-statistic to evaluate the effect at the sample level. The t values of adjacent points with the threshold of p value < 0.05 are clustered together, and a cluster-mass statistics is calculated by summing the t values within a cluster. Once these candidate clusters are obtained, we would like to estimate how big clusters would be under the null hypothesis of no difference between the conditions. To assess their significance, a nonparametric permutation test is used, in which conditions are shuffled and cluster-mass t values are calculated as before. This step is repeated 2000 times and on each iteration the largest cluster t value is retained. From the distribution of these summed t-values, the significance of the observed candidate clusters is calculated as the proportion of expected t values that are more extreme than the observed ones (see [Maris and Oostenveld, 2007] for further details). Our main interest in this analysis was primarily motivated from the opportunity it provides to check the neural activity across all 64 electrodes with no pre-defined region of interest, with no priority of finding significant clusters.

4.4.1.2. Slicing and rotating the vowel wheel

As a preliminary investigation of periodicity in the neural data, and to understand the changes in the EEG signal as a function of F1 and F2, for the values of the periodicity *n* between 1 and 6, we divided the trajectories into two conditions, because the way our clustering method is implemented works only with 2 conditions. For example, for 1-fold symmetry, the data of 30 trajectories are divided into two sets each containing 15 trajectories, and clustering analysis is applied. For n = 1, this is repeated for 15 rotations of the wheel to check for every possible combination of the trajectories in 2 main clusters of directions. In our example, what the clustering analysis would help us to find is if there is one particular direction that the neural activity is significantly sensitive to, so that there is a significant difference reflected in the amplitude of the neural activity in the processing of the wide band trajectories starting and ending in opposite directions, and by which angle that divides the vowel plane this difference is maximized, in addition to when and in which region it occurs in the brain. The same reasoning is used to perform the analysis with larger values of the periodicity. This is somewhat an indirect analysis of the periodicity, but useful to get an idea about the signal, as there are many possibilities to check.

We inspected all the clusters obtained with an arbitrary threshold of $p < 0.3$. The time window of the analysis was from the onset of the stimulus till 180ms after end of it, but we mainly focused on the clusters that show some difference around 100ms and 200ms, as our earlier finding in a separate study suggested a modulation of the activity based on the first two formants around these two time points, and on the clusters that last longer than 5ms.

We were not able to find such ‘meaningful’ clusters for many of the conditions we explored. The activity of one of the promising clusters (t = -113.26, p = 0.2389) we observed is shown in Fig. 4.15 for the 6-fold symmetry condition, as illustrated on the small wheel overlaid on the same figure. A maximal difference between two conditions, which are blue and red trajectories at every 5^{th} direction as depicted on the wheel, is reached around 85ms stimulus onset latency over the 10 right occipito-parietal electrodes shown in the topography map of the difference.

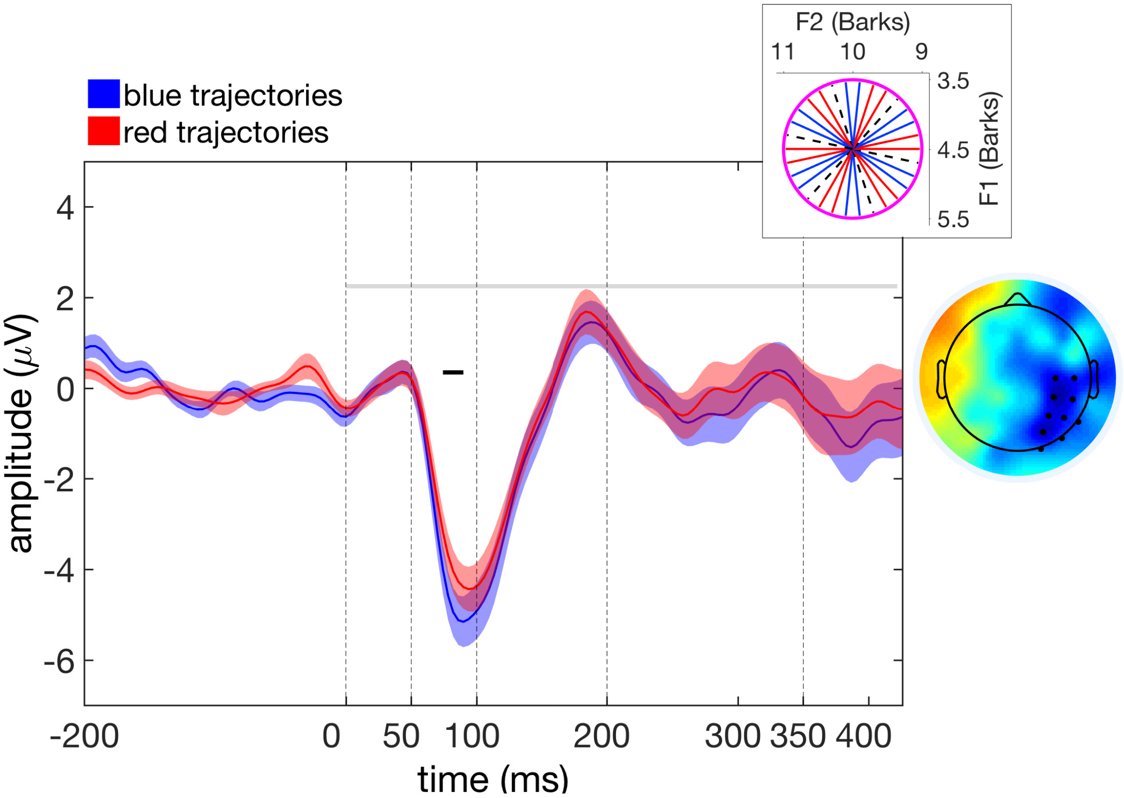


Fig. 4.15: Grand average of two set of trajectories at every 5^{th} direction. Time 0 indicates the onset of the stimulus. Error bands denote the SEM across participants. The horizontal light grey line delimits time the window of interest. The horizontal black line delimits the time range of the cluster (p > 0.05). The top inset shows the trajectories in 2 conditions at every 5 directions. (Right) Topography of the difference wave between the 2 conditions averaged across the time the cluster is observed.

In our next analysis, we used these 10 electrodes to search for periodicity directly. As a control, we also included 10 left occipito-parietal electrodes, that locate on the mirror symmetrical positions of the ones above, and a set of 10 fronto-central electrodes, a region of interest in the studies of auditory processing [Picton, 1974, Bruneau et al., 1997, Khalighinejad et al., 2017]

4.4.2. Fourier components of the EEG signal

For every subject, trajectory, and electrode, we first averaged the EEG data over trials. 2 participants with the data of 1 trajectory missing, and 1 participant with the data of 2 trajectories missing due to the cleaning in the preprocessing were discarded in this analysis. Therefore, the reported results are from the data of the remaining 19 participants.

Before the Fourier decomposition of the signal done, as explained in Section 4.3.1, we normalized the signal by turning it into a z-score, that is calculated relative to a standard deviation that is different for each subject, channel, and time bin across 30 trajectories, to make the comparisons easier. The z-score for each participant at each time bin is calculated as:

$$z(\theta) = \frac{r(\theta) - < r >}{\sigma} \tag{4.5}$$

where θ is the trajectory angle, <r> is the average signal amplitude across trajectories, and σ is the standard deviation of $r(\theta)$ along the wheel.

Then from the signal averaged across the three regions of interest, we identified four extrema: a peak at about 45ms (P50), followed by a deep plunge with a bottom around 85ms (N100), followed again by a peak around 190ms (P200), and a final positive peak around 340ms (P350) as also shown in the example of Fig. 4.15 (see the vertical dashed lines in light grey).

We looked at the periodicity the signal exhibits at these extrema individually, but in order to clean potential variability among the individual participants that lowers the signal-to-noise ratio, we computed the difference including [P50-N100], [P200-N100], [P350-N100], and some other for each group of electrodes separately for each subject. In the left panel of Fig. 4.16, in three different regions, we show the power spectrum, i.e. the squared magnitude of the Fourier Series coefficients (i.e. weights), averaged across subjects where smaller colored circles at each fold denote the data of individual participants. The signal of each subject is taken as the difference between P50 and N100. The increase after 11-fold in left occipito-parietal and fronto-central electrodes give an impression of the high noise level in the signal. The neural activity in the right occipito-parietal electrodes show a promising 7-fold symmetry, but looking at the data at subject level, as we did in the right panel of Fig. 4.16, we see that is mostly due to very few, i.e. 2, subjects. In the same panel, the power spectrum of the subtracted EEG signal of one participant displaying low 7-th component is shown with dashed lines, as an example.

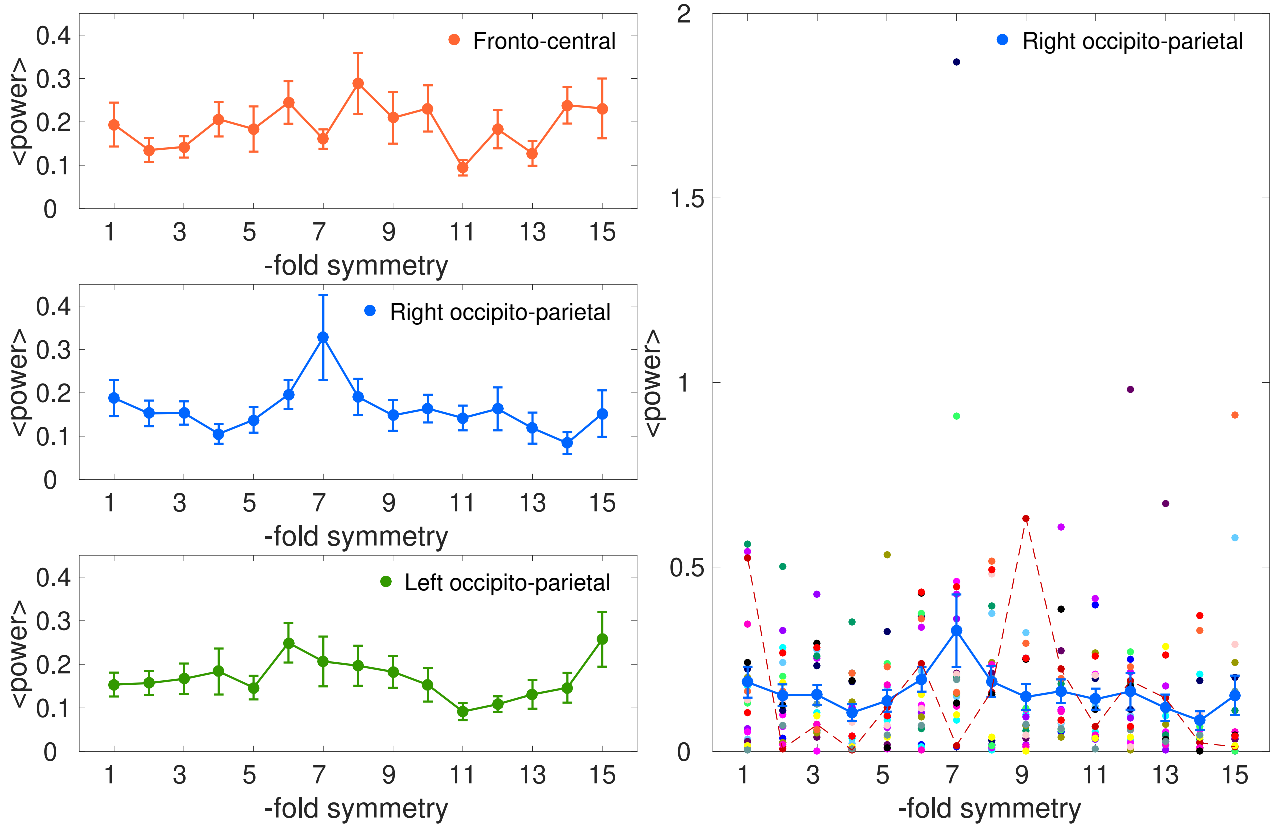


Fig. 4.16: Mean power spectrum of the EEG signal in three different regions. (Left) The EEG signal computed as the difference between P50 and N100 does not display a particular periodicity in any 3 regions of interest, as seen from the average power spectrum. (Right) 10 electrodes in the right occipito-parietal region show a greater 7-fold symmetry compared to the other components, as also seen in the middle figure of the left panel, but the individual data points of 19 participants in colored circles show that it is mostly due to 2 subjects with high 7-fold coefficient, hence the large error bars demoting SEM across participants. The dashed lines show the power spectrum of the EEG signal of an exemplar subject that has low 7th component.

In the same manner, not to limit our analysis to a specific time point such as [P50-N100] as in Fig. 4.16, we extended our analysis to the whole time course of one epoch, that starts 200ms before the sound onset and ends 185ms after the end of the sound, by subtracting N100 peak as shown in Fig. 4.17. However, we did not observe periodic behavior at other different time points of the signal either. We think the 7-th component we observe at around 50ms, is too soon and too short to be considered as a real effect of the navigation.

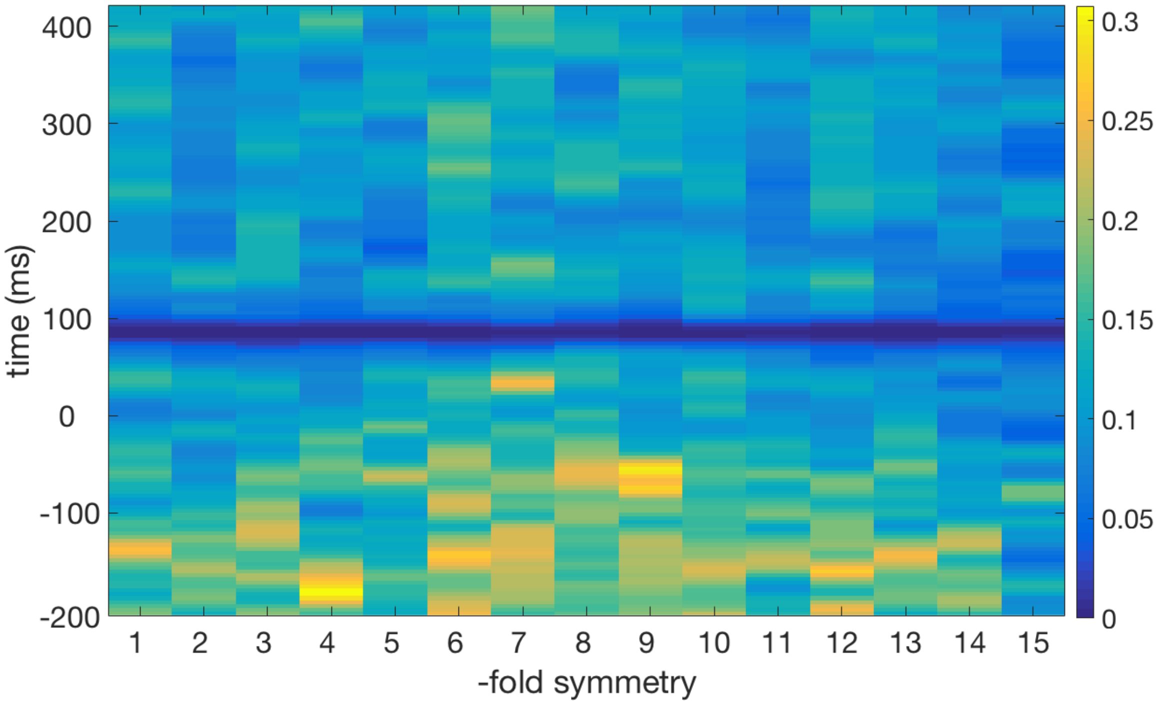


Fig. 4.17: Mean power spectrum of the EEG signal of the right occipito-parietal electrodes along the time course of an epoch. The substracted EEG signal shows that 7-fold symmetry is not persistent through time. But rather almost all of the Fourier coefficients play an equal role in the signal structure after stimulus onset at 0ms.

We also looked for the periodicity in another set of electrodes that are between left and right occipito-parietal regions at around 240ms, due to our observation of another cluster ($p > 0.05$) with the analysis in Section 4.4.1.2, but also there, we could not find a significant Fourier component.

4.5. Modulation of the signal along the wheel

4.5.1. Can we compare outside the center to inside?

We put the wheel in the center of the vowel space, where there are no standard vowel categories in Italian. Vowels along the wheel may be thought to act like landmarks that guide navigation in spatial environments, and through extensive training, the mental representation of this central space might become flattened (in Bark coordinates). Do we see any interesting modulation of the neural signal according to the first two formant frequencies along the wheel, in this empty center?

Outside of the center, the evoked potentials of native Italian listeners to (American) English consonant-vowel pairs, which end with one of the four standard vowels /e/ /a/ /i/ /u/, suggest two levels of processing of the formant frequencies, which occur at two different time points (see Chapter 2.11 for the experimental paradigm). As seen in the top plot of Fig. 4.18, around 100ms (N100), the neural trajectories are clustered into two groups based on the degree of frontness, or backness of the ending vowels. The EEG signal of the front vowels /e/ and /i/ shows a greater negative deflection than the back vowels /a/ and /u/. In the same figure, we see, around 200 ms (P200) after syllable onset, the grouping of the neural signal is based on the degree of openness, and that the amplitude of the positive deflection for the open vowels /e/ and /a/ is greater than that for the close vowels /i/ and /u/. Is this mechanism of F2 and F1 processing as reflected by the N100 and P200 components also valid for the neural processing of the empty central region?

In order to see if such an imprint of formant processing is also observed for the neural correlates of our quasi-diphthongs, we looked at the EEG signal of the trajectories that belong to different portions of the wheel. We grouped 3 consecutive trajectories on the wheel into 5 groups of different diphthongs, (not all equi-spaced), according to the proximity of their starting direction to 5 standard vowels, /e/ /a/ /i/ /u/ and /o/ as in the inset of the bottom plot in Fig. 4.18 (see also Fig. 4.6 for the positions of standard vowels in British English space).

The bottom plot of Fig. 4.18 shows that at around 100ms, although the processing of the steady initial components of the four diphthongs, /e/ /a/ /i/ and /u/, is reminiscent of the processing of the vowels of the outer region, the neural signal of the steady part that starts as /o/ casts a doubt on the the model which assumes F2 processing at N100. Given our observations on the vowel processing in the perimeter of the vowel space, we would expect the neural correlates of /o/ to be similar to those of /a/ and /u/, as it is a back vowel.

However, we should also remember that what participants hear at 100ms in the two experiments is not the same. In the first experiment with CV syllables, what they hear is a steady vowel (or a transition into it from a consonant). That is different from what they hear in the current experiment

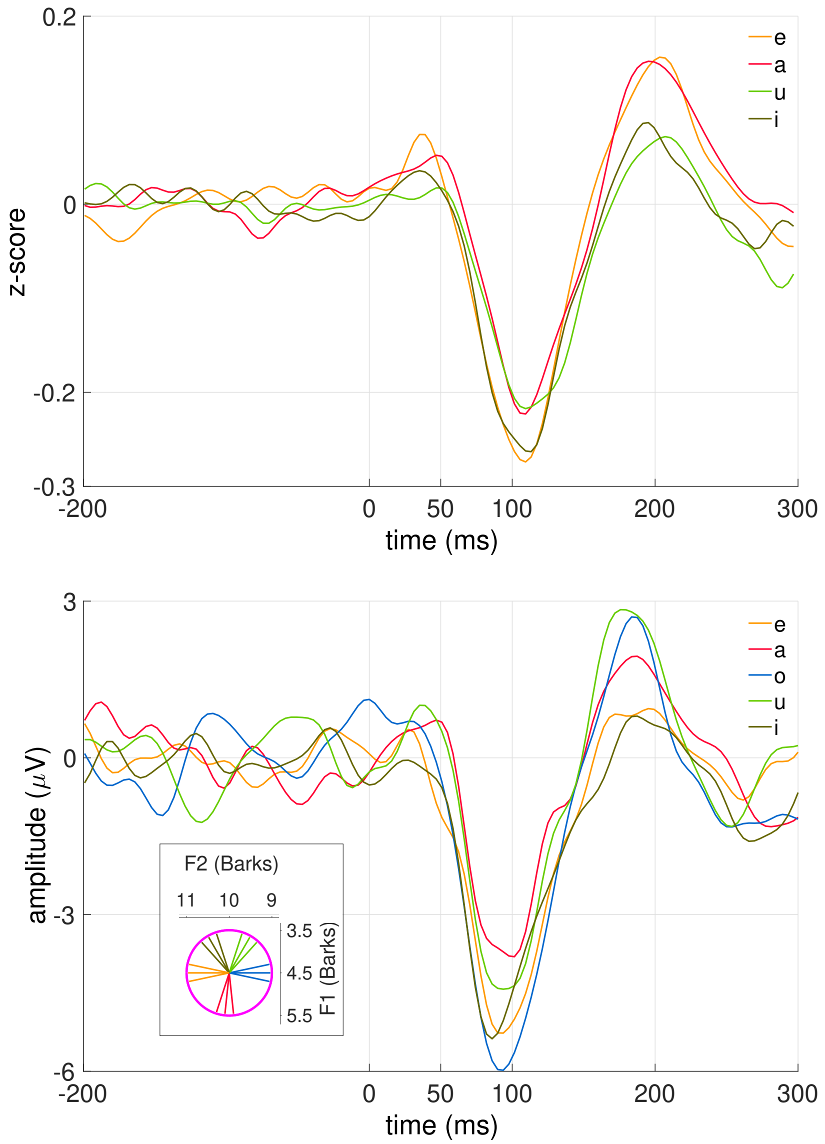


Fig. 4.18: Grand averages of vowel related potentials in and out of the central region. (Top) Z-scored evoked potentials of CV syllables ending in /e/,/a/,/u/, and /i/ averaged across 64 electrodes (see Chapter 2.11). (Bottom) Evoked potentials of diphthongs on the wheel that start from the points close to five standard vowels (see the inset) averaged across 64 electrodes. The processing of the vowels of the outer region suggested F2 processing in N100 and F1 processing in P200 components. However, in the center, the trajectory with the deepest N100, that is a trajectory that starts as an /o/, which is in the roughly 180° opposite direction to the one that starts as an /e/, contradicts the pattern observed for the vowels outside of the center. The P200 component reflects even stronger differences between the center and the perimeter.

we are discussing. In addition to the 20ms long steady portion, there have been the initial 80ms of the long dynamic part of the acoustic signal that has F1 and/or F2 changing in the same or opposite directions (see the inset of Fig. 4.18 and also Fig. 4.7). It is puzzling to see that although /e/ and /o/ start from opposite ends of the F2 axis, the difference between the EEG amplitudes of their perception is similar, but of opposite sign, to the difference between the perception of /i/ and /u/ trajectories that have closer second formants. Also puzzling, the difference between the amplitudes of the perception of /i/ and /e/ is much smaller than the difference between /o/ and /u/ even if the F2 distance between /e/ and /i/ is similar, in both cases small, as the one between /o/ and /u/.

Moreover, at P200, there is no corresponding clustering between the two panels of Fig. 4.18 Looking at the bottom plot in the same figure, we see that the trajectories that start as /o/ or /u/ have similar amplitudes. Likewise the trajectories that start as /e/ or /i/ have also similar amplitudes. In between these two clusters, there are the trajectories that start as /a/. Is this reflecting an F2 processing, which we observed 100ms earlier with steady vowels? Again, by 200ms, what participants hear is a 20 ms steady part and a longer portion of the dynamic of the acoustic trajectory. For example, for /o/ and /e/ trajectories, the participants hear trajectories with the same mean F1 and F2 for 160ms but changing in two opposite directions. However, the difference between their signal is even amplified compared to their difference at 100ms.

What is the difference between the two neural imprints in the center and outside the center due to? Can it be because the sounds are squeezed in a small space at the center, too close to each other, with the same mean formant frequency? Can it be because they are not stable in time and what N100 and P200 reflect includes the dynamic part of the acoustic signal? Can the variability in the EEG signal across trajectories inform us about the neural representation of the wheel?

4.5.2. What does variability in the signal tell about the central region?

In order to understand the variation in the signal as a function of trajectory direction, we performed a linear transformation of the neural data by principal component analysis. We first averaged the EEG signal across 3 individual electrodes among 10 in each of the 3 regions (left occipito-parietal, fronto-central, and right occipito-parietal), and then averaged across all subjects. For every trajectory, we looked at the signal of individual subjects calculated from the difference between 3 extrema and N100, which are [P50-N100], [P200-N100], and [P350-N100] in addition to the difference between the temporal mean of the signal and N100, that is [Temporal mean - N100], to suppress individual variability.

To center the data at zero, for each of the 4 mentioned differences between two peaks, and between the N100 peak and the average across time, we subtracted the mean of the signal across trajectories at those points for every subject instead of z-scoring the ERPs along the wheel. Finally, to repress the variability that is due to the limited number of trials per trajectory, we smoothed the ERPs with a quasi-Gaussian approximation, which is calculated for every trajectory as:

$$r_g(\theta) = 0.05 * r(\theta - 24^\circ) + 0.25 * r(\theta - 12^\circ) + 0.40 * r(\theta) + 0.25 * r(\theta + 12^\circ) + 0.05 * r(\theta + 24^\circ) \tag{4.6}$$

where $r(\theta)$ is the value of the ERP for the trajectory angle, which we denote with θ.

The final data is denoted by a 12-by-1 column vector S, where S_{ij} represents the signal as a function of trajectory direction which correspond to the i-th region at the j-th time point:

$$S = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ \vdots \\ S_{12} \end{bmatrix} = \begin{bmatrix} S_{1,1} \\ S_{1,2} \\ S_{1,3} \\ S_{1,4} \\ \vdots \\ S_{3,4} \end{bmatrix} \tag{4.7}$$

Then the 12-by-12 covariance matrix $\sum$ is the matrix whose (i,j) entry is the covariance:

$$\sum_{i,j} = cov(S_i, S_j) = \frac{1}{30} \sum_{\theta=1}^{30} (S_{i,\theta} - E(S_i))(S_{j,\theta} - E(S_j)) \tag{4.8}$$

where $E(S_i)$ and $E(S_j)$ are the averages across trajectories of the i-th and j-th entry in the vector S.

Using the covariance matrix, we can obtain 12 eigenvectors of size 12-by-1 whose eigenvalues are given in Table 4.1. Looking at the eigenvalues, we see a strong dominance of the first one, which accounts for 61% of the variance. A sharp decrease after the first eigenmode steadily continues till the 6th eigenvalue, after which only 1% of the variance is accounted by the successive eigenvalues.

	Eigenvalues	
3 electrodes	1 electrode	10 electrodes
2.8642	3.1551	2.6752
0.8298	0.8493	0.5895
0.2999	0.4136	0.2259
0.2581	0.3567	0.2803
0.1624	0.2036	0.1298
0.1361	0.148	0.1103
0.0436	0.0636	0.0446
0.0297	0.0346	0.0217
0.0113	0.0158	0.0096
0.0086	0.0114	0.0062
0.0012	0.0015	0.001
0.0005	0.0005	0.0006

Table 4.1: Strong dominance of the first eigenmode. When 3 electrodes per region are considered, among the 12 eigenvalues obtained from the covariance matrix, there is a strong dominance of the first eigenmode, that is 3 and half times greater than the second, and 9 times the third. After the first 6 eigenvalues, there is a drop to less than 1% of the variance for the successive ones. When only 1 electrode is chosen, there is a trade-off between an increase in the first eigenmode and an increase in variance (not shown; so that their ratio slightly decreases with respect to the 3 electrodes case). When all the electrodes in a region are taken into account, the first eigenmode is slightly decreased.

In Fig.4.19 and in Fig. 4.20, for each eigenvector V_i, we plot the linear sum W_i of the selected time difference signal weighted with the i-th eigenvector, that is:

$$W_i = \sum_{j=1}^{12} V_{ij} S_j \tag{4.9}$$

The error for each trajectory has been calculated as the SEM across participants:

$$Error_\theta = \sqrt{\frac{E(W_{i,\theta}^2) - E(W_{i,\theta})^2}{N}} \tag{4.10}$$

where $E(W_{i,\theta})$ is the mean across participants at a given trajectory, and N is the number of

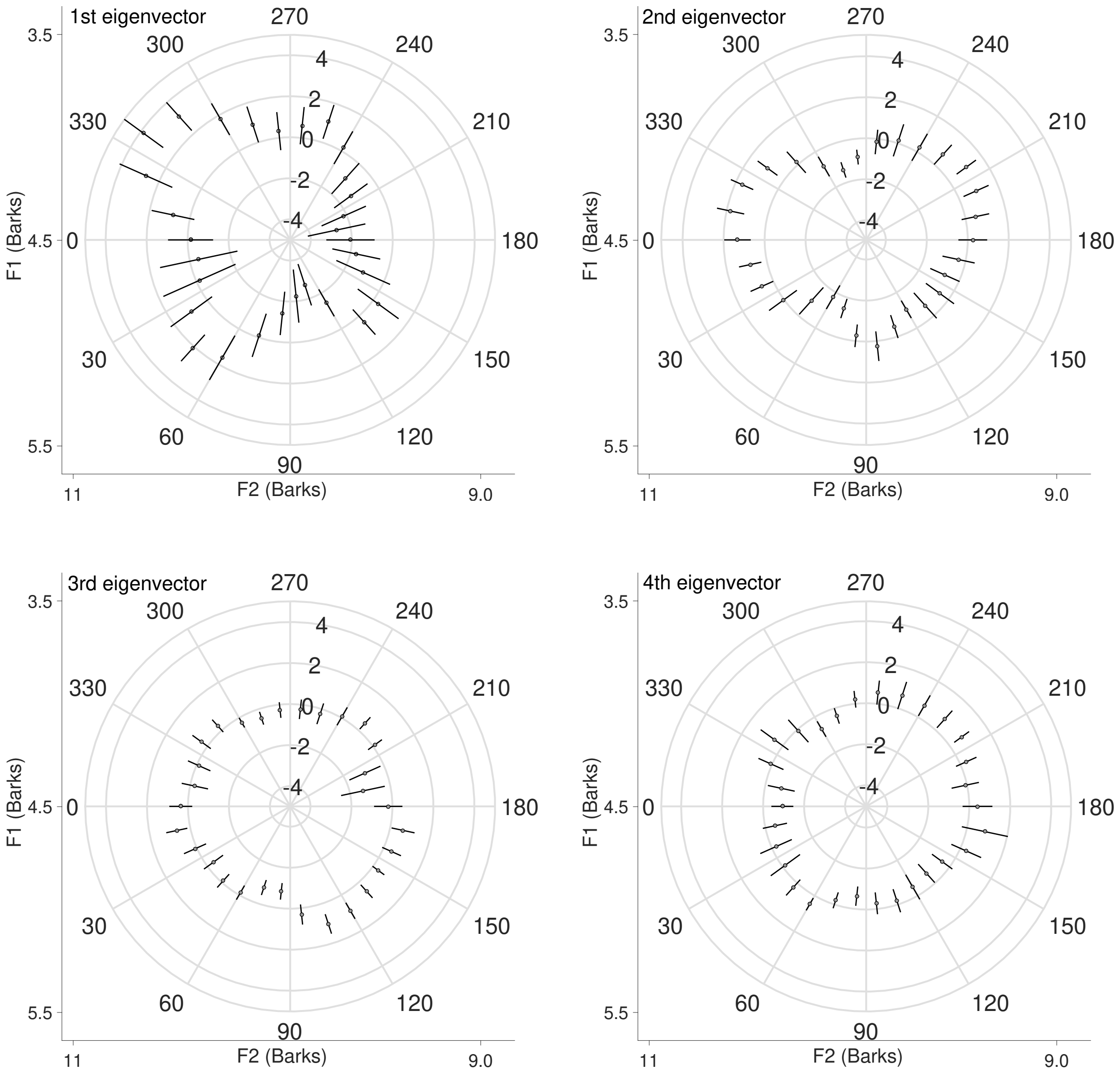


Fig. 4.19: EEG signal across 30 trajectories weighted with the first 4 eigenvectors seperately. Each polar plot shows the linear combination of the EEG signal with the first (top left), second (top right), third (bottom left), and forth (bottom right) eigenvectors respectively. The minimum and maximum radius limits are restricted to the mean of the weighted amplitude ± 5. The signal weighted with the first principal component shows the greatest deformation along the 4 main diagonal directions on the wheel, but mostly for the ones which start between 300° and 330°. For the second eigenmode, the deformation is in two main directions along F2, and for the third eigenmode, the deformation is mostly in the direction of trajectories starting between 90° and 180°. The variability vanished already when weighted with the 4th eigenvector.

participants, 19.

The shape of each polar plot in Fig. 4.19 gives us an idea about the modulation of the EEG signal along the wheel, where the limits of the each radius axis are adjusted to the mean of the particular weighted amplitude ± 5. Since the first eigenmode is the strongest one, i.e. the one that explains the most of the variance in the EEG signal, the signal transformed into the direction of the first eigenvector is more informative about the change in the signal as a function of direction, as seen on the top left polar plot in Fig. 4.19. The first eigenmode is roughly excited along the 4 main diagonal directions, which start between 30° and 60°, between 120° and 150°, between 240° and 270°, but mostly between 300° and 330°, that is the region where the trajectories start close to an /i/ sound, and end close to an /o/ or ɔ.

For the other eigenmodes, the deformations on the wheel appear to occur along fewer directions and, in particular for the second eigenmode, they seem to be restricted to two main directions along F2, and for the third eigenmode, a bulge is mostly in the direction of trajectories starting between 90° and 180°, where sounds start with low F2 and high F1, and end with high F2 and low F1. These deformations, shaped by the majority of the variance in the data, do not relate to the performance which is strongly peaked between 240° and 300° (see Fig. 4.11).

The signal has already a quasi-circular shape when weighted with the fourth eigenvector (see bottom right in Fig. 4.19), and looking at the weighted sum of the signal with the other 8 eigenvalues does not lead to any informative observation, as they account for less than 1% of the variance in the data. As seen in Fig. 4.20, the shape of the signal is more and more of a circle when weighted by the eigenvectors between 7 and 12.

We repeated the same analysis considering the average of all 10 electrodes in each region, and also taking only one electrode per region of interest. In Table 4.1, we show the respective eigenvalues for different choices of number of electrodes to be considered for the EEG signal. When considered averaging across three aforementioned regions across all of their 10 electrodes, there is a slight decrease in the first eigenmode to 2.68, with a similar shape of the eigenvector and variability. On the other hand, working with just one electrode, although resulted in an increased first eigenmode, that is 3.16, the calculated mean error also increased due to a higher level of noise arising from the variability among subjects, so that the ratio of eigenvalue to variance actually decreased. Thus, the choice of 3 electrodes per region appears a reasonable one, as it is probably

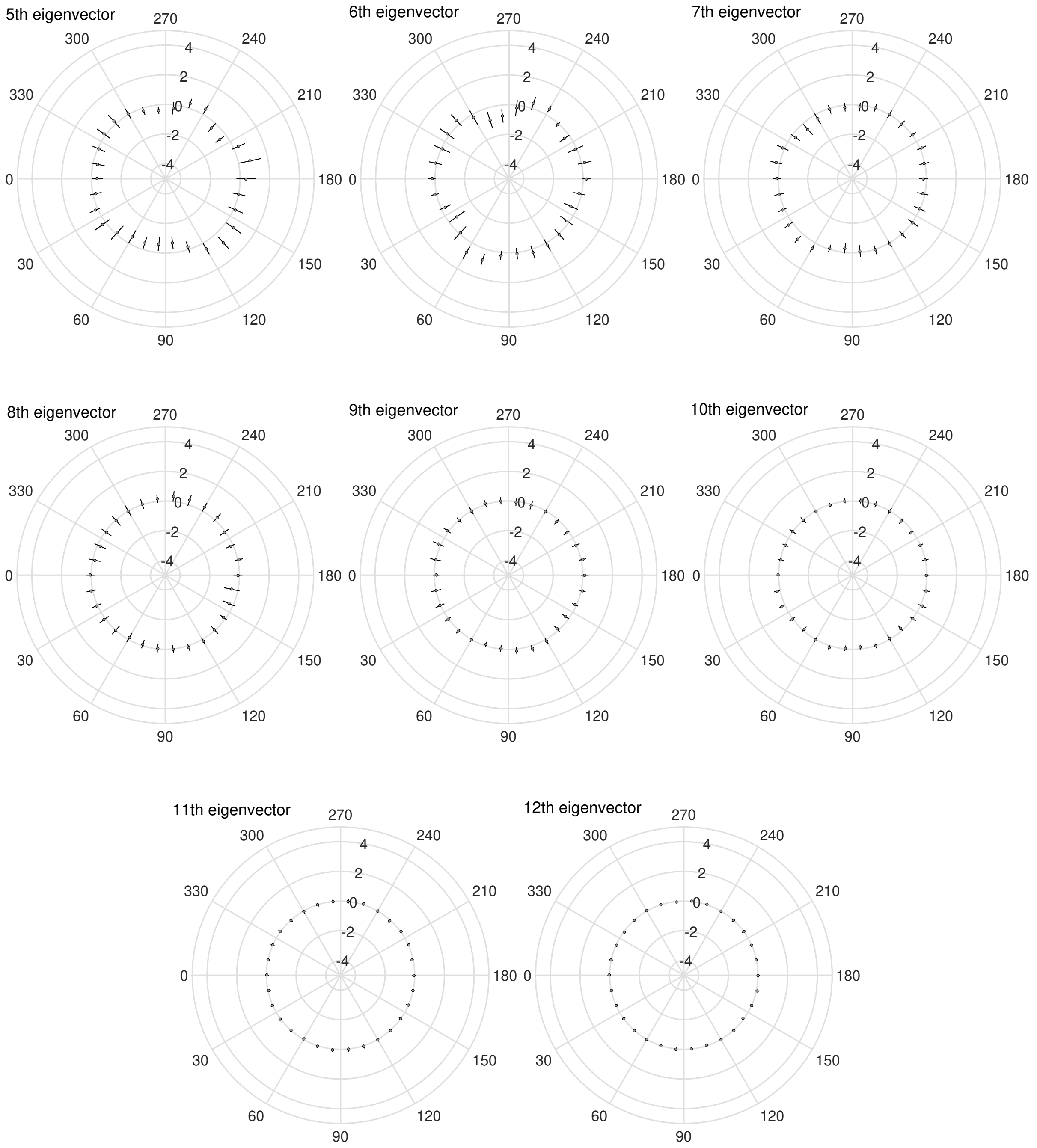


Fig. 4.20: EEG signal across 30 trajectories weighted with the last 8 eigenvectors seperately. Each polar plot shows the linear combination of the EEG signal with the last 8 eigenvectors. As in Fig. 4.19, the minimum and maximum radius limits are restricted to the mean of the weighted amplitude ± 5. Compared to the first two figures on top, staring from the 7th eigenvector (top rightmost) till the 12th (bottom rightmost), the shape of the signal along the wheel becomes more and more circular as the last 6 eigenvectors account for negligible fractions of the variance.

close to optimal in terms of signal to noise ratio. Obviously taking somewhat more or fewer electrodes per region does not change results too much (data not shown). This is confirmed by the Pearson correlation, which is calculated between the linear combinations of the ERPs along the wheel when different number of electrodes are considered, as following:

$$\rho_{W_{i,x},W_{i,y}} = \frac{cov(W_{i,x}, W_{i,y})}{\sigma_{W_{i,x}}\sigma_{W_{i,y}}} \tag{4.11}$$

where $W_{i,x}$ and $W_{i,y}$ refer to the weighted sum of the EEG signal with the same eigenvector i for different number of electrodes x and y. cov and σ refer to covariance and standard deviation respectively. Pearson correlation coefficients are given in the columns of Table 4.2 which shows that the results of PCA remain similar regardless of the numbers of electrodes we used.

	1 electrode & 3 electrodes	3 electrodes & 10 electrodes
n-th Eigenvector		
1^{st}	0.991	0.995
2^{nd}	0.957	0.942
3^{rd}	0.836	0.941
4^{th}	0.784	0.933
5^{th}	0.843	0.883
6^{th}	0.89	0.923

Table 4.2: Pearson correlation between the weighted EEG signal when averaged over different number of electrodes. When EEG signal is averaged over different number of electrodes and projected along the direction of different principal components, its shape along the wheel remains the same as revealed by the high correlation coefficients.

4.5.3. The last positive peak

Which components of first eigenvector play a more important role in the variation of EEG signal? In other words, which time points among the EEG extrema vary the most along the wheel?

We observed that the weight of the fourth component of every region, which corresponds to the difference between the last positive peak and the first negative peak, i.e. [P350 - N100], is greater than the difference between other time points and N100, on the first eigenvector. This means that in order to vary the most along the wheel one should move toward the direction which is mainly parallel to the [P350 - N100]. Therefore, the shape of the principal component, as shown in Fig. 4.19, resembles the shape of the EEG signal of the forth component, as seen in Fig. 4.21 separately for each region.

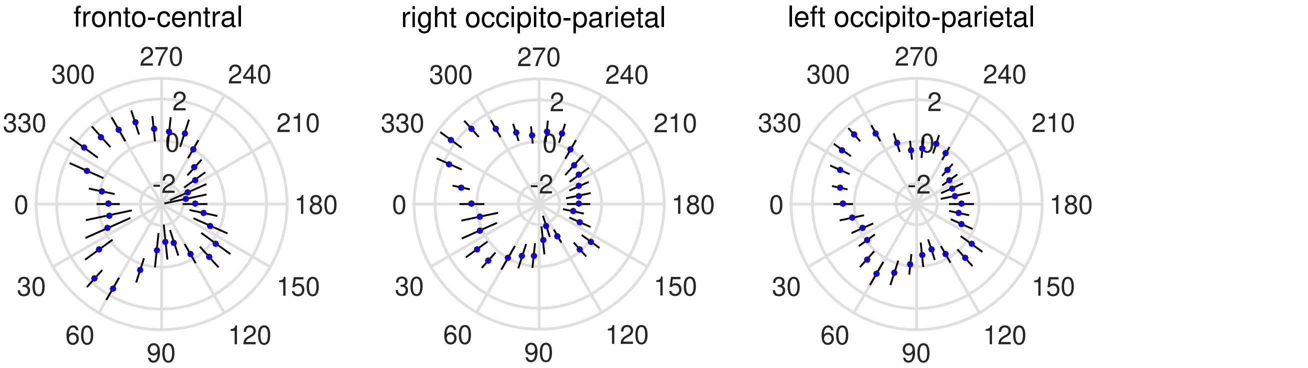


Fig. 4.21: EEG signal of [P350-N100] in three regions. The first eigenvector (see Fig. 4.19 has a shape highly similar to the subtracted EEG signal between P350 and N100, which dominates it in every region, especially fronto-central and right occipito-parietal regions. Left occipito-parietal region has a more inflated and less edgy shape.

How is the neural processing around the last positive peak relevant and why does it have a strong dominance compared to the processing at earlier time points? We do not exactly know, but looking at the grand-averages of correct and wrong trials of all sounds across the 10 electrodes in the fronto-central and right occipito-parietal regions, as in Fig. 4.22, especially for the former, we see the difference between the two, although small, is greater in the last positive peak, but also in the last negative peak. Interestingly, while in the fronto-central region, the deflection for wrong trials is bigger than for the correct ones, it is reversed from what is exhibited by the right occipito-parietal region. In Fig. 4.22 we do not show left occipito-parietal electrodes, but we should note that this difference occurring at the last positive peak is particularly reduced there. Overall, the neural population activity around this last peak might be an important component of the perception and thus of the final decision.

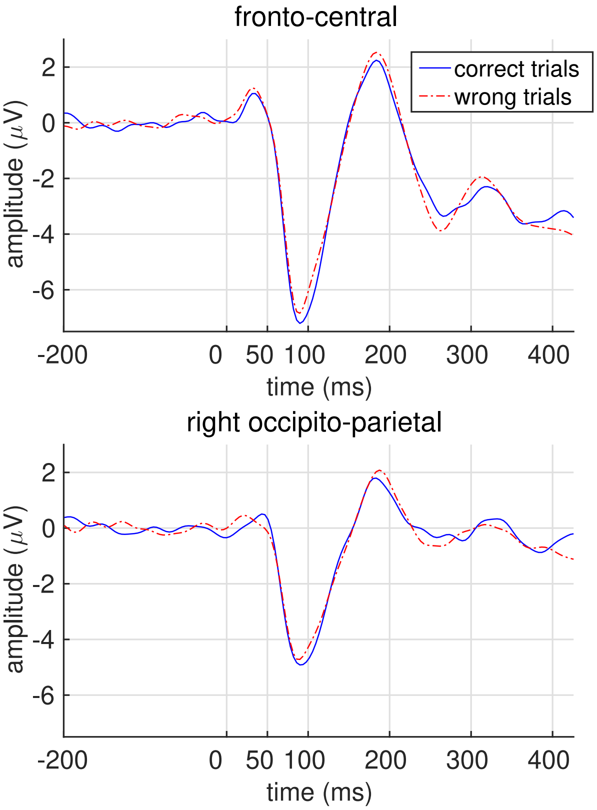


Fig. 4.22: Grand average of all correct and wrong trials in two regions. The last negative and positive peaks, before and after 300ms, show the maximum difference between two conditions in both fronto-central and right occipito-parietal electrodes. For the latter, the last peak shows a bigger positive deflection for correct trials, while for the former it is reversed.

4.6. Discussion

It is tempting to speculate that we represent this overly familiar portion of the (F1, F2) plane somewhere in our brain in similar ways to how we, and other mammals, represent familiar 2D spatial environments. A lot is known, in particular, about the representations rodents acquire of boxes or small arenas where they are left to forage for food, in terms of place cells, grid cells and other varieties of neuronal activity [O'Keefe and Dostrovsky, 1971], [Fhyn et al., 2008], [Taube et al., 1990], [Solstad et al., 2008], [Kropff et al., 2015]. Grid cells compromise a Eucledian metric in empty arenas [Moser and Moser, 2008], but we could not detect a neural signal of phoneme trajectory perception expressing hexagonal symmetry in the center of the 2D vowel plane.

One possible reason could be that the vowel space with radius of 1 Bark is not large enough, after all, to develop a grid like representation. Even when it is navigated coast-to-coast, so to speak, in a /io/ diphthong for example, this is done in 100-200 ms in natural speech - possibly too short a time for any unit even inclined to be a grid cell to express multiple fields.

Still the shape of the first eigenvector suggests that mass activity is not isotropic (rotationally invariant) around the wheel. Is it because neural activity even inside the wheel is sensitive to the presence outside of the standard vowels? That would be a natural outcome of the vowel space being "too small". But if so, is this influence language dependent, as it should be if it reflects standard vowels which are language dependent? We test this hypothesis, that the space is small, hence dominated by the standard vowels, and hence language dependent, by analyzing behavior in the next chapter.

One could wonder whether there is a relation of the first eigenmode to existing vowel attractors. We designed the wheel to be in the empty space in the middle among standard Italian vowels. Although the position of standard vowels in any language is highly variable and speaker and context dependent, we can refer e.g. to the authoritative early study by Ferro et al. and express the formant frequencies they cite in Barks [Ferro et al., 1978], and place them around our wheel. Even taking other reference values, the standard vowels are clearly outside the wheel, as shown in 4.23.

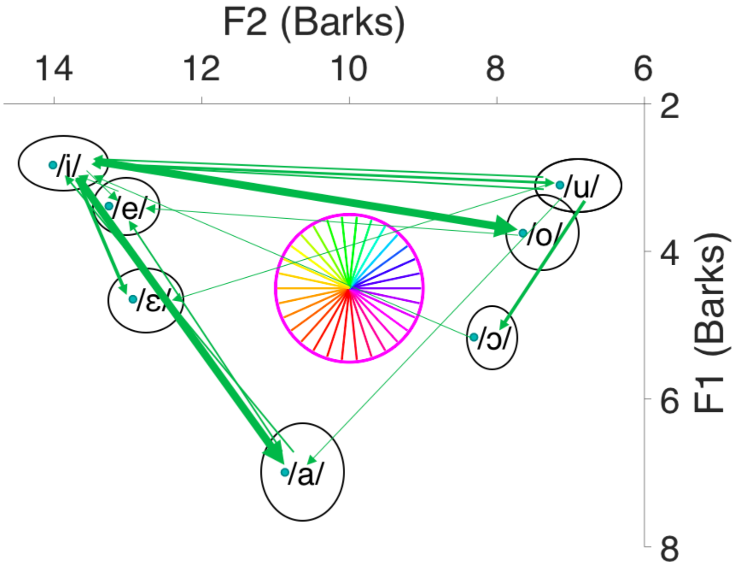


Fig. 4.23: Diphthong trajectories in the Italian vowel space. Standard Italian vowels are placed on the Bark space according to the formant frequencies described in [Ferro et al., 1978]. The lines denote the Italian diphthongs widths of which are proportional to the frequency of usage [Goslin et al., 2014]. The vowel wheel we used in the experiment is in a relatively untouched region.

Further, even common Italian diphthongs do not appear to interact much with the wheel. In Fig. 4.23, between the standard vowels, we show approximate diphthong trajectories, with the thickness of the lines proportional to the frequency of usage of the diphthong or rather biphone combinations in the Phonitalia database [Goslin et al., 2014]. The thick lines, mainly /ia/ or /ja/, /io/, also /jɛ/, /ju/, /wɔ/, are all outside the wheel. Therefore we can assume that our wheel trajectories, although diphthong-like, are quite distinct from real common Italian diphthongs and their putative attractors.

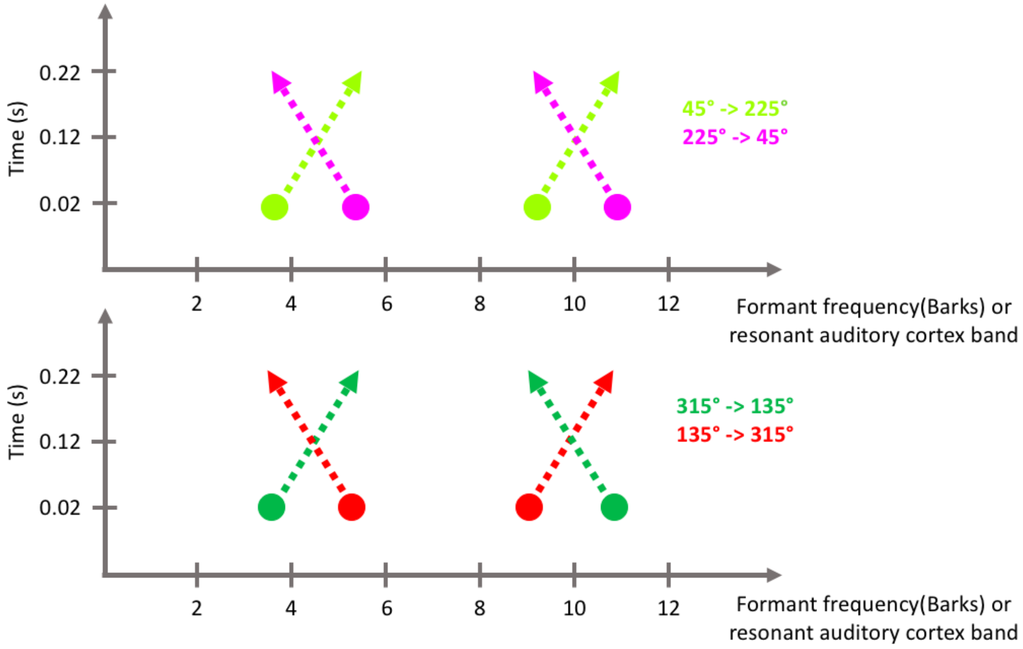


Fig. 4.24: An illustration of the physical properties of the most excited diagonal trajectories. Can the shape of the first principle component of the EEG signal be related to the topographic organization of auditory cortex by spectral frequency? (Top) The first and the second formants of the two trajectories (45°-225° in light green) both decrease together. The 180° opposite trajectories (in pink) both increase in parallel. (Bottom) The trajectories on the most excited diagonal (315°-135° in dark green) have increasing first formant and decreasing second formant. The trajectories 180° across (in red) have the opposite relationship.

What else could the dominating eigenmode be related to? One possibility can be inferred from the fact that the eigenmode is excited most close to the 4 diagonal trajectories on the wheel. The two 45°-225°(/ɛ/-/a/ to /u/-/w/) trajectories are those where the first and second formant rise or fall in parallel. In auditory cortex there is tonotopy, with a metric not known in detail, but which we can assume to be close to a Bark scale. Therefore while the fundamental frequency will be close to constant in time along our trajectories, the first and second resonant frequencies will be two bands of excited neurons moving in parallel one direction or the other (top row in Fig. 4.24). The most excited diagonal however is the one going from 315°to 135°(/i/-/y/ to /a/-/ɔ/), which corresponds to two coliding bands of excitation. Interestingly, the opposite direction is relatively less excited, the red one, corresponding to two diverging bands of excitation. Maybe these physical traits give rise to the eigenmode, which in this case would be language independent, probably. We leave it for a future study to present the wheel test to subject speaking other languages, and so test whether their first eigenvector is different from that for Italians.

5

Vowel Segments

5.1. Defining one's own metric in vowel space

Phoneme perception, like many other linguistic processes, is thought to involve discrete variables, which in the case of phonemes have to be parsed from the auditory input stream. In many instances, especially with consonants as we discussed in Chapter 2, the phonological variables reflect features which can be taken to be already discrete when the sounds are produced, for example nasalization or voicing. In contrast, as we pointed out several times, vowel space is inherently continuous (see, for example Chapter 1).

The vowels of natural languages can be described to a good approximation as mapped onto a continuous 2D Euclidean space, a plane whose axes correspond to the frequencies of the first and second vowel formants, F1 and F2. While for some languages some additional features are necessary to specify a vowel, such as its length or roundedness (see Chapter 3), whose nature may be more discrete, for many others F1 and F2 are all there is to it. F1 and F2 simply reflect the extent to which the vocal tract is open, and where the main occlusion lies, during vowel production. Their physiological range, in Hertz, can be converted into a perceptual scale, e.g. in Barks, through a simple one-to-one transformation, initially linear and then logarithmic [Zwicker, 1961] (see Chapter 4.2.1) .

Although individual languages effectively employ this space as if it were divided into a small number of quasi-discrete basins (i.e. the vowel inventory of a given language), the space is intrinsically continuous. Note that while languages vary from one another both in terms of how they divide the acoustic space, and how the divisions are treated by the grammar, the outer boundaries of the space itself are relatively invariant cross-linguistically, owing to consistent physiology between speakers of different languages (barring individual differences in vocal tract length, etc.) [Peterson and Barney, 1952]. This is in keeping with Ferdinand de Saussure's observation that the relation between sound and meaning in human language is arbitrary [de Saussure, 1966]. For example, a sound exactly half-way between the words 'hat' and 'hut' cannot point to a meaning half-way between these words, but rather must be parsed as meaning one, the other, or neither. And yet, the basins of attraction of individual vowels cannot be too rigidly associated to regions on the vowel plane, if anything because different speakers utter them differently, one's *hat* can be another one's *hut*, and listeners have to maintain a speaker-specific frame of reference floating over the plane, as it were, in order to parse vowels correctly [Fox, 1982,Lehiste and Meltzer, 1973].

In the previous chapter we investigated a small and untouched portion of the vowel space in which we expected, with some extra help, vowel trajectory perception to reflect a regular grid metric, as it is the region most distant from any possible attractor, and therefore to find a hexagonal modulation of the underlying neural signal. However, the empty space with a radius of 1 Bark was perhaps too small for the grid expression to emerge. What about the larger region covering outside the center? Grid cells tend to express a Euclidian metric in extended empty arenas [Moser and Moser, 2008], but that metric can be distorted by the boundaries [Stensola et al., 2015]. We can ask, therefore, whether our representation of the vowel space, in its linguistically relevant portion, is distorted by the presence of 'objects', the phonemes of our vowel repertoire.

Given that at least two main aspects of each vowel sound (F1 and F2) can be regarded simply as the coordinates of a point on a 2D Euclidean space, do native speakers of different languages, who use different standard vowels, perceive the similarity between sounds in agreement with their physical Euclidean distance, or with a more general perceptual distance, largely shared among speakers of the same language but different for speakers of languages with significantly different vowel systems? We test the hypothesis with an experiment, wherein we assess to what extent native speakers of different languages, whose vowel inventories differ, perceive the similarity between sounds not in agreement with their ‘physical’ Euclidean distance (measured in Barks), but rather with a language specific perceptual distance.

5.2. Materials and methods

5.2.1. Stimuli

We have created a morph continuum between two vowels in a pair. There were four vowel pairs in total: /u/-/y/, /o/-/œ/, /æ/-/ɛ/, and /ɔ/-/ʌ/. Fig. 5.1 shows the positions of the two end points of each continuum on the vowel diagram. All vowels were pronounced twice; once preceded by a voiced and once by a voiceless version of a consonant by a British male speaker. All (4*2 = 8) stimuli [/vu/, /fu/, /vy/, /fy/, /bo/, /po/, /bœ/, /pœ/, /dæ/, /tæ/, /de/, /te/, /gɔ/, /kɔ/, /gʌ/, /kʌ/] were normalized to the same RMS using Praat [Boersma and Weenik, 2018].

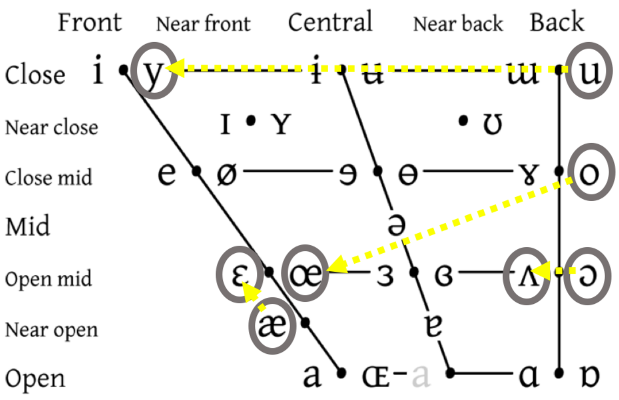


Fig. 5.1: The vowel segments shown on the standard vowel chart. Grey circles are imposed on the vowel categories between which we created morphing continua, as represented by yellow dashed lines. Some segments are made up of closer sounds, hence the dashed lines are shorter, and some others are made up of distant categories and passing through some other categories that exist in some languages.

We used the STRAIGHT library running under Matlab [Kawahara et al., 1999] to obtain two intermediate points between two syllables starting with the same consonant, for example between /fu/ and /fy/. STRAIGHT extracts from the two waveforms several parameters, aligns them temporally, and then interpolates between the two according to the anchor points. The anchor points, which consisted of the first two formants and temporal limits of the vowel, were chosen carefully with a phonetician. The two intermediate points had a morph rate of 0.33 and 0.66 in order. One exemplar vowel continuum is shown in Fig. 5.2. If we can denote the four points in a segment by AAA, AAZ, AZZ and ZZZ, as in Fig. 5.2 and if AAA is /u/ and ZZZ is /y/, then AAZ is 2/3 /u/ and 1/3 /y/, AZZ is 1/3 /u/ and 2/3 /y/. All 4*2*4 (32) stimuli were trimmed to 400ms.

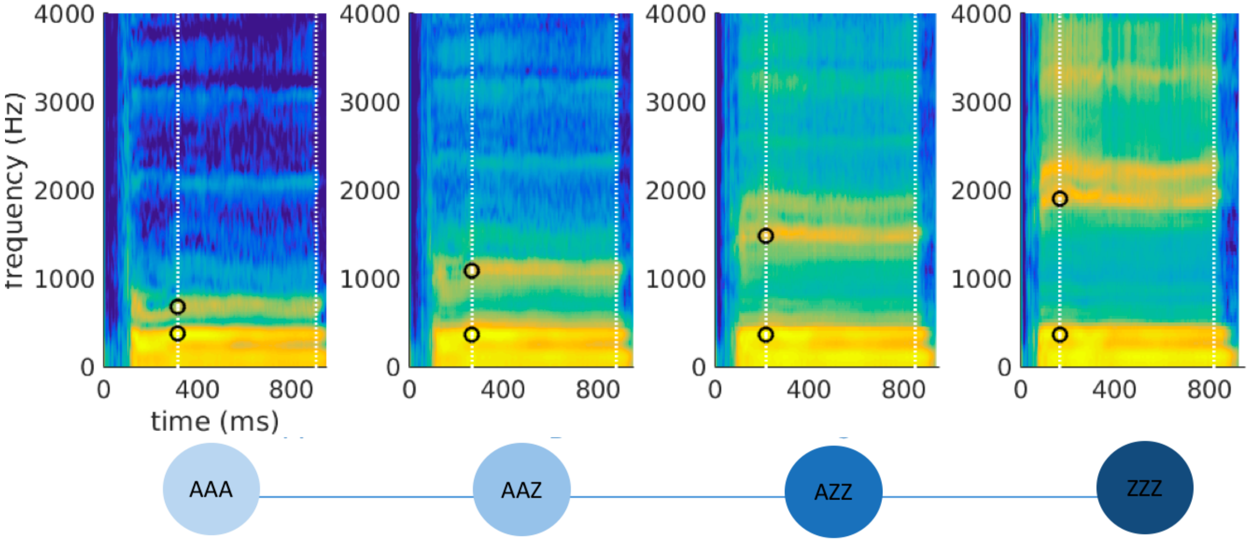


Fig. 5.2: The vowel continuum created between /fu/ and /fy/. Two end points AAA and ZZZ are /fu/ and /fy/ syllables pronounced by a British male speaker. We created two intermediate points, AAZ and AZZ, on this continuum, with morphing rates of 0.33 and 0.66. The morphing algorithm makes use of time-frequency landmarks to define correspondence between two vowels. The selected spectral anchor points, i.e. the first and second formants, are plotted as open circles, and the selected temporal anchors, i.e. the onset and offset of the vowel, are plotted as vertical dash-dot lines.

5.2.2. Participants

We have tested native speakers of Italian, (non Catalan/Valencian) Spanish, Turkish, and Scottish English. In each language group, there were 16 participants and all the participants were non-bilingual and could speak English, but perhaps with varying degrees of their competence. The vowel space of each language, which differ in their category and number of vowels, can be seen in Fig. 5.3. Spanish has the least number of vowels, and both Spanish and Italian have their vowels sitting closer to the edges, while Turkish and Scottish English have richer set of monophthongs, with some extra vowels in the inner region.

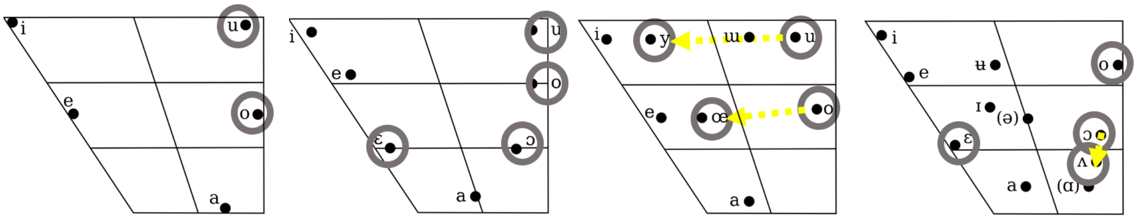


Fig. 5.3: Vowel diagrams of 4 language groups. From left to right, we see the vowel charts of Spanish, Italian, Turkish, and Scottish English. Spanish has only 5 vowel categories. Both Spanish and Italian have their vowels on the periphery of the diagram. In comparison, the charts of Turkish and Scottish is more crowded, and some of their vowels lay in the inside region of the trapezoids. Some accents of Scottish pronunciation have the central schwa sound denoted inside parentheses. Vowel categories and segments included in the experiment are denoted by circles and yellow dashed lines on each plane, if the categories are recognized as standard by speakers of that language.

5.2.3. Paradigm

32 (4*8) utterances were rearranged into 4*40 (160) S1-S2 pairs, where S1 and S2 were always the two distinct voiced and voiceless variants of the consonant, to minimize confounds. Among 40 trials included for each vowel segment, S1 was the same as S2 in 16 trials, or adjacent to S2 in 12 trials, or S1 was two apart from S2 in 8 trials, or three apart in 4 trials. An illustration of the presented syllable pairs according to the distance between them is shown in Fig. 5.4.

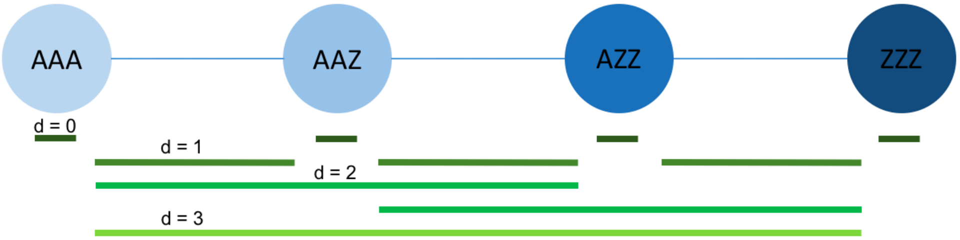


Fig. 5.4: Experimental paradigm. Pairs of syllables with different distances between the two are presented to the participants. For each one of the four vowel pairs, there were 40 trials. 16 of these trials consisted of the same two syllables (with distance = 0). In the rest of the 24 trials, the syllables had distance of 1 (in 12 trials), 2 (in 8 trials), or 3 (in 4 trials). Figs. 5.5 and 5.6 may be regarded as looking at this diagram from one side, while Fig. 5.7 as looking at it from below.

For every contrast, the vowel sounds in S1 and S2, 2 presentations had S1 with the voiced and the other 2 with the voiceless consonant, and S2 with the opposite one. The presentation order of S1 and S2 was swapped so that half of the trials started with S1 and half of them with S2.

160 trials were randomized so that no two consecutive trials would be of the same vowel segment to avoid short-term memory effects. Every trial started with 1500ms of silence, followed by the presentation of two 400ms long syllables with a 300ms long silence in between them. The participants were asked to respond if they perceived the two sounds as 'same' by pressing the space bar in 2000ms, after which a new trial started unless they already made a response before. For the analysis, we accepted the late responses up to 100ms after 2 seconds long response limit as 'same'.

5.3. Different memory representations at work?

As discussed earlier in Section 2.6, adult speakers of different languages are expected to express different analog-to-digital recording of these sounds, which are not all in their native language. How are continuous vowel dimensions F1 and F2 related to discrete sound categories? Comparing perceptual differences among the different parts of the same vowel segment, and the same segments across subjects with different vowel repertoires can help us understand the encoding of vowels in long term memory. How are these operations influenced by the basins of attraction formed by the mother tongue?

The nature of representation of a particular vowel segment can be continuous and even linear in terms of the physical properties, i.e. the first and second formants, characterizing vowel space. In that case, the response to ambiguous stimuli (AAZ and AZZ in Fig. 5.4), those that are a mixture of different sounds, would be a linear interpolation of the un-morphed end-points (AAA and ZZZ in Fig. 5.4), as reflected in the proportion of different responses in our behavioral experiment. A representation of this kind is illustrated in Fig. 5.5 with the dashed lines.

On the other hand, perception of sound identity may have a categorical nature, involving less gradual and more abrupt transitions in the proportion of different responses across the morph continuum. Two different examples of this kind of representation are shown by the pink and purple lines in Fig. 5.5. In such scenarios, the perception of intermediate morphs would be close either to one or to both of the end points. This has been argued to be the case with

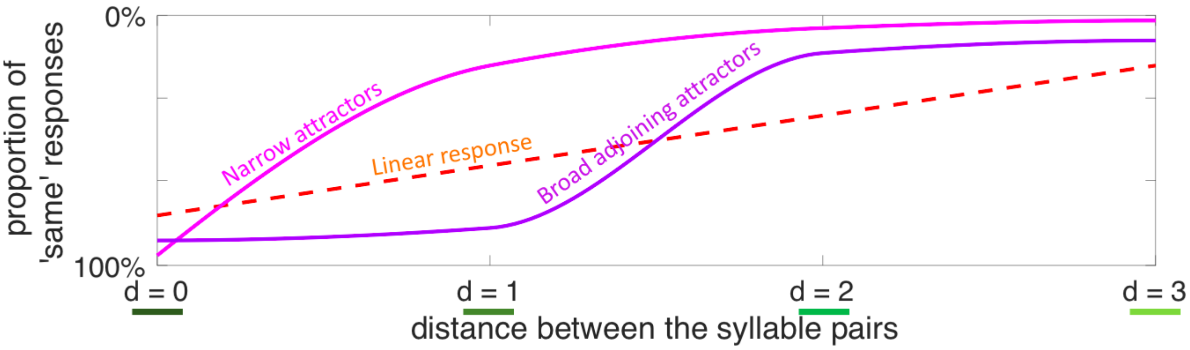


Fig. 5.5: Possible scenarios of discretization of continuous F1 and F2 dimensions. What is the relationship between the distance of two sounds and the perception as distinct categories? Looking at the proportion of 'same' responses among trials consisting of syllables with different distances (as in Fig. 5.4), we can tell if perception depends linearly on the distance or if it is categorical and reflects basins of attraction.

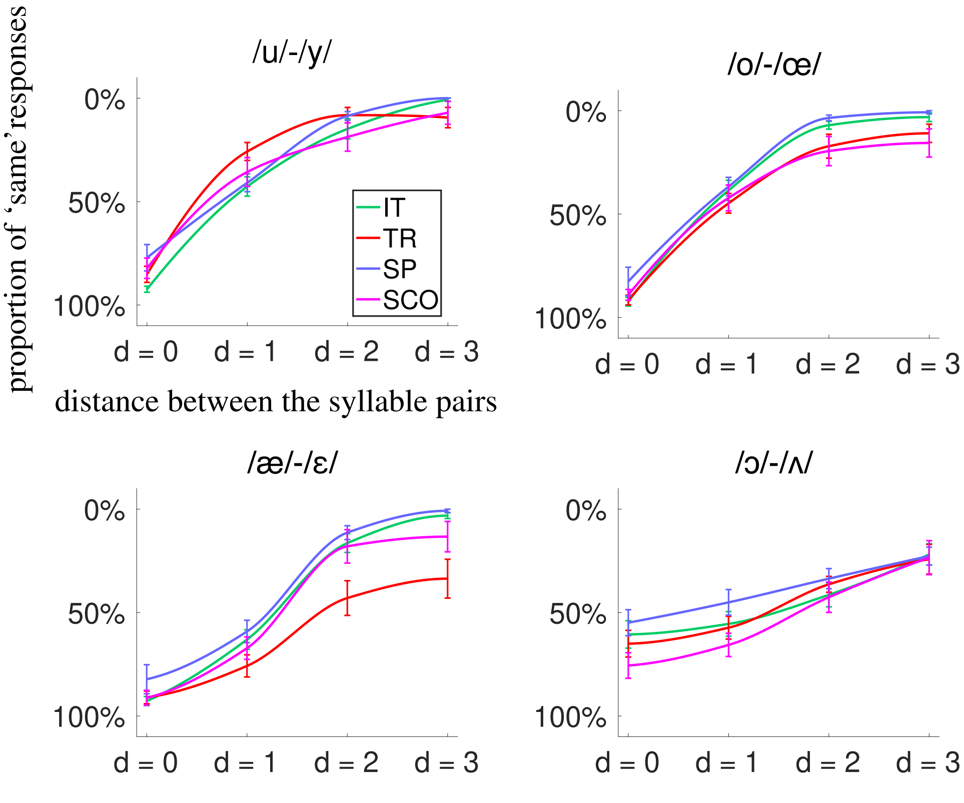


Fig. 5.6: Processing of different vowel segments by 4 different mother tongue groups. Psychophysics curves show that although different languages show broadly similar patterns for each vowel segment, the processing of different vowel segments are not exactly the same. They differ in their linearity, with the segment made up of the two nearby categories /ɔ/- /ʌ/ (see Fig. 5.1) being processed most linearly compared to the other three segments. The /æ/-/ɛ/ segment appears closest to the 'broad adjoining attractors' type, likely because it spans a limited range of formant frequencies, hence its putative end-point attractors are pushed towards each other.

consonants [Liberman et al., 1957], but what about vowels, the production of which takes place in a continuous manifold?

As we see from the psychophysics curves in Fig. 5.6, where we plot the proportion of ‘same’ responses among pairs grouped according to the distance between them, each vowel segment is processed differently, although different mother tongue speakers show broadly similar patterns. Among four segments, the segment formed by two nearby vowels /ɔ/ and /ʌ/ is the one that is processed the most linearly. Even in the two end points, every language we tested, including Scottish that has both categories, seem to have difficulty in discriminating the sounds, presumably due to larger difference between the voiced and unvoiced syllables, orthogonal to the effect of the frequency comparison they make along the segment. For the segment /æ/-/ɛ/, in which the two end points are also closer to each other than the sounds in the other pairs /u/-/y/ and /o/-/œ/ (see Fig. 5.1), both vowels seem to form attractors, but it is less pronounced for Turkish speakers possibly because they lack these sounds. For the other two segments, it seems that the attractors are narrow (relative to the significant length of the segments), so most of the ‘same’ responses are concentrated at d = 0.

This ‘side view’ on the segment represented in the diagram of Fig. 5.4 suggests differences between vowel perception by speakers of different languages, but these differences may be masked by averaging, in the analysis above, over the discrimination of sounds that are at the same ’distance’ along the morphing dimension. In the analysis below, we take a ‘frontal view’ instead, in order to examine effects at different location on the continua.

5.4. Local Metrics

Another way of showing the four local metrics we obtained is to plot the frequency of ‘same’ responses in a ’frontal’ view, without averaging over pairs of sounds at the same distance. In Fig. 5.7, the mean same responses for syllable pairs made up of the same sound are shown with the thinnest and darkest bars, and the same responses for vowel pairs made up of two different sounds are the bars that have the width and hue corresponding to the distance between two sounds.

We can make several observations on these attractor bars that inform us about the local geometry of the plane. For example, we see that for Turkish speakers, the connection between the morphs is

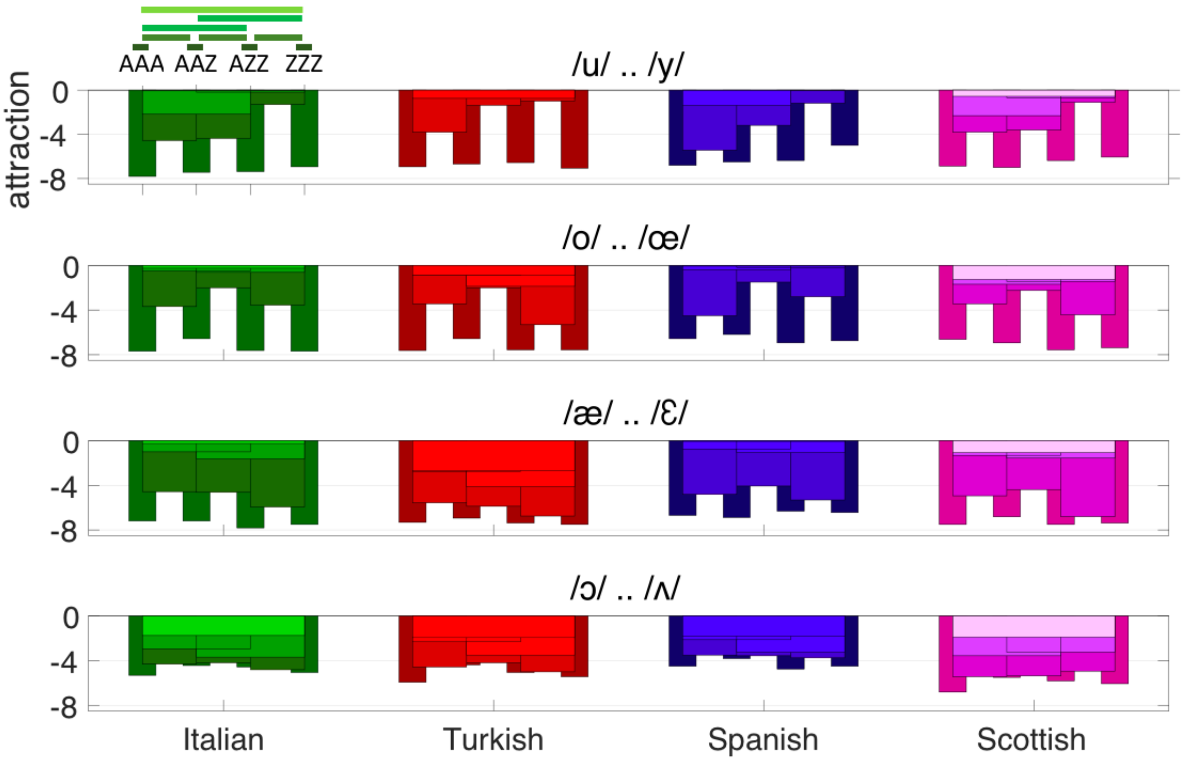


Fig. 5.7: Local metrics of different languages. The mean frequency of same responses of syllable pairs of same sounds (/AAA/-/AAA/, /AAZ/-/AAZ/,/AZZ/-/AZZ/,/ZZZ/-/ZZZ/) is shown with the thinnest downward attractor bars. The frequency of same responses to syllable pairs of different sounds are the bars that have a width corresponding to the distance between the sounds. These metrics inform us about local deformations of the vowel plane that are specific to each language.

not as deep for the /u/-/y/ continuum as it in the other languages. This might because in Turkish there is another vowel category /ɯ/ laying on this segment. However, in the three other languages, in addition to the deeper bar between the second and third morph, there exits the wide connection between the first and the third points that does not occur in Turkish. Another observation we make is on the perception by Turkish speakers of the /æ/-/ɛ/ segment, where all the bars are connected to each other with thick bridges of the larger widths. Their discrimination of the trajectories on this segment is worse as they do not have these two sounds, but only nearby /e/. For the /o/-/œ/ segment, the connection between the third point and the last is deeper than or equal to the one between the first and the second for all the language groups except Spanish that shows the reverse trend. Finally, the /ɔ/ and /ʌ/ segment is highly interconnected, across language groups.

The panels of Fig. 5.7 indicate some differences between the language groups, as we have discussed, but overall a remarkable similarity, expressed for example in the relative isolation of the /y/ sound from the rest of its segment, across all groups, or in the split in two of the /o/-/oe/ segment. Is this due to the local nature of the analyses afforded by the 4 morph segments? To address this question, and consider more global analyses, it is necessary to first return to the issue of a global metric valid throughout the vowel plane, as already mentioned in Section 5.1.

5.5. Defining a global metric

Let us display the morph segment results on a (presumed) metric representation of the entire vowel plane.

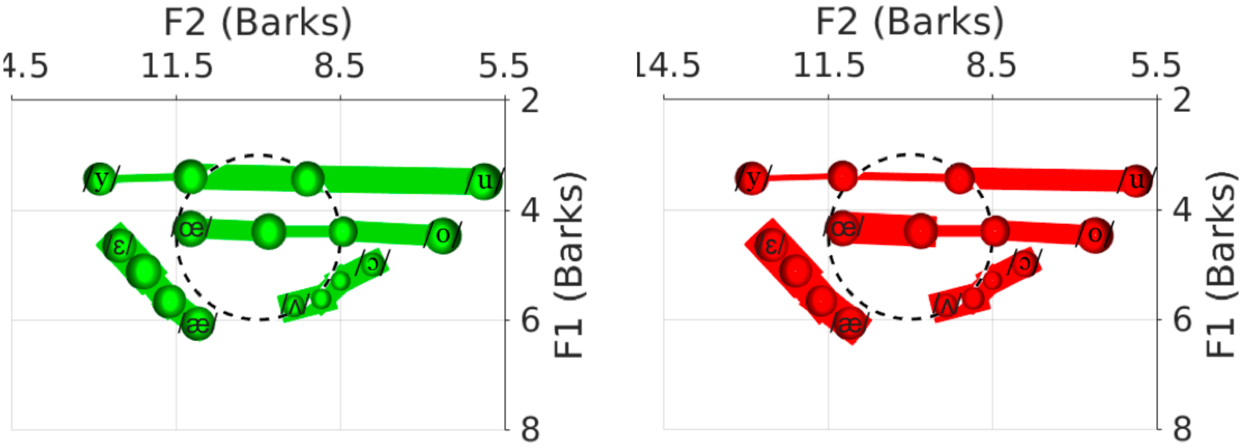


Fig. 5.8: Perceptual vowel segments of native Italian and Turkish speakers. The vowels are laid on a grid in the Bark Scale. Turkish (in red on the right) and Italian (in green on the left) vowel spaces differ mainly in the inner central region, delimited by the dashed circle here. While the circle is empty in Italian vowel space, it has two additional vowels /w/ and /œ/ in Turkish.

In the top and the bottom panels of Fig. 4.2.1, the sounds we used in that experiment are placed on a cartesian grid in a Bark Scale, in which equal steps in formants correspond to approximately equal steps of perceptual distance, as explained in Section 4.2.1. The coordinates in Bark are obtained by the average of the formants in the voiced and voiceless syllables, which adds jitter to the exact positions in addition to the nonlinearity of the transformation from Hertz to Barks, and hence some sounds are off the theoretical intermediate points. The radius of the spheres and the width of the channels are proportional to the probability of identifying the vowels in a pair as same by Italian speakers on the top panel, and by Turkish speakers on the bottom panel. We do not show the perceptual spaces of Spanish and Scottish speakers as they look very similar to the one of native Italian speakers.

The central region that is empty in languages like Italian and Spanish is shown with the dashed-line circle with a radius of 1.5 Barks. This is a slightly bigger circle than the one we used in the experiment of Chapter 4, which was of radius 1 Bark, to be more safely inside the empty region. This region is however not empty, for example not in Turkish, but also not in Scottish (see Fig. 5.3). Inside the circle, Turkish has the vowels /ɯ/ in the upper part and /œ/ in the middle part. As explained previously in Section 5.4, the portion between the two morphs on /u/-/y/ segment, where /ɯ/ lays, tends to be identified less as same, and hence the thinner belt between them compared to the same in Italian perceptual space. A similar differentiation between Turkish and Italian (or the other two languages) is not observed in the middle part of the circle. The distances between /œ/ and the morph closer to it, and between that morph and the other morph closer to /o/ are approximately the same, but perception of these two portions are not equal in Turkish and in Italian. As /œ/ belongs to the vowel inventory of Turkish, the first portion is perceived slightly more similar in Turkish than it is in Italian, and in the other two languages, but the difference is smaller with Scottish, which does not have the vowel /œ/ but other sounds close-by.

Given the similarities and differences between the perceptual spaces as they are reflected in the central region, we see that the responses inside the vowel plane do not reflect a regular grid even in languages where the center is empty. Earlier, in Chapter 4, we hypothesized that this region could be flattened in languages where there are no standard vowel categories, and we used extensive training, with trajectories, i.e. quasi-diphthongs to obtain a regular grid metric. Perceptually we managed to achieve a relatively uniform representation of this space, but what is the nature of the representation as it is, when tested with point-like sounds, i.e. single vowels? Is this representation language-dependent?

5.6. The triangulation experiment

In order to understand the specific transformation that applies to each language, to deform the representation from a square metric, in our next experiment we covered the entire space by adding new morphs and contrasts, i.e edges, as shown in the left panel of Fig. 5.9.

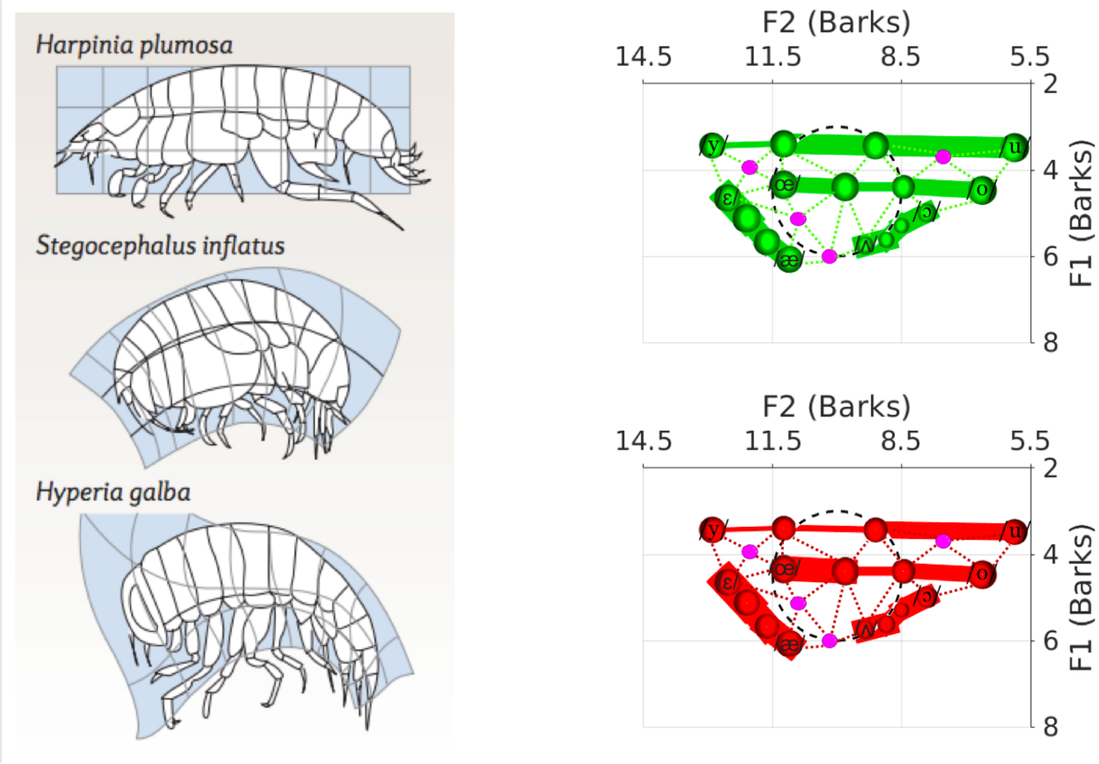


Fig. 5.9: Language-specific transformation of the vowel space. (Left) D'Arcy Thompson transformed a simple creature he represented in a cartesian grid (top figure) to find its phylogenetic relatives (middle and bottom figures), as here shown for different crustaceans (Original figure taken from [Wallace, 2006]). (Right) Can we find a language-specific transformation similar to D'Arcy Thompson's, if we connect all the segments of the previous analysis with four extra vowels (pink spheres)?

This is similar to what we see on the left panel of the same figure. The three figures on the left are the drawings by Scottish biologist and mathematician D'Arcy Thompson, who would find phylogenetic relatives by representing an initially simpler creature on a grid (top most) and then applying a mathematical transformation, such as stretching in one dimension or distorting, which would result in a shape that would be of another animal related to the transformed one (middle and bottom) [Wallace, 2006, Thompson, 1942].

5.7. Materials and methods

5.7.1. Stimuli

All the sounds we used can ben seen, in Bark coordinates, in Fig. 5.10. The bark coordinates of each sound is the average of its coordinates in voiceless and voiced syllables, which adds jitter to the exact positions in addition to the nonlinearity of the transformation from Hertz to Barks. We used the same sounds of the previous experiment (see Section 5.2.1) and using the same procedure as in the previous experiment, we created 4 additional morphs, in the figure denoted by number 8, 7, 6, and 12, to connect the vowel segments.

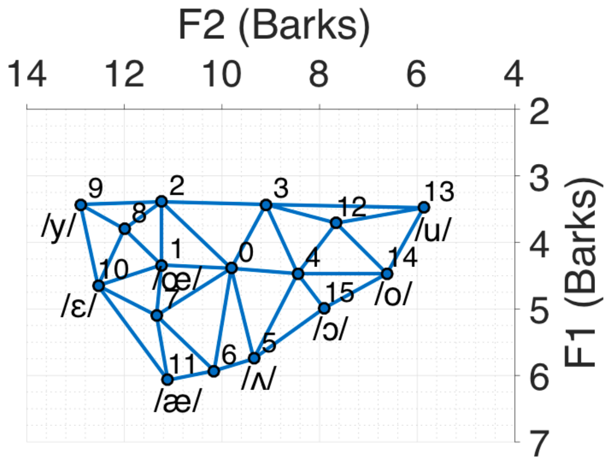


Fig. 5.10: The physical space of the sounds used in the experiment. The vowel segments used in the previous experiment are stitched together to each other with the four new morphs by the new contrasts, i.e. edges, presented to the participants. There are a total of 35 edges in the triangulation shown.

The new morphs were designed to lay on the intersection of two different pairs of two sounds, except the sound 6. For this reason, for every new morph, we first obtained a morph from the two ends of one pair with a morphing rate of 0.5, and repeated the same for the other pair. This resulted in two morph candidates per one voicing contrast. By generating a new morph from these two candidates with the same rate, we produced the final morph to be used in the experiment. For example, in order to generate the sound 8, we first created a morph between 9 and 1, and another one between 10 and 2. In order to average these two sounds, we created a morph between them that is the sound 8. The sound 7, which is in the middle between 5 sounds, was obtained by the crossings of the sounds 10-6 and 1-11. The sound 6 was obtained by morphing only between the sounds 11 and 5.

5.7.2. Participants

We have tested native speakers of Italian, Turkish, and Norwegian. In each language group, there were 20 participants and all the participants were non-bilingual and could speak English, but perhaps with varying degrees of their competence. The rich vowel space of (Standard East or Urban East) Norwegian can be compared to the vowel spaces of Italian and Turkish, which we previously discussed in Fig. 5.11. There are many dialects in Norway, and officially there is no standard spoken language. However, dialects with richer vowel inventories than the East Norwegian urban standard tend to reduce their vowel system and match the one shown here [Kaun, 2010].

Among the eight vowels we used in the experiment to generate the morphs, Norwegian possesses six of them. The symbols /e/ and /ɛ/, and /ø/ and /œ/ are used interchangeably in different sources on Norwegian phonology [Kristoffersen, 2000, Vanvik, 1985]. In contrast to Italian and Turkish, most Norwegian speakers can discriminate the schwa /ə/.

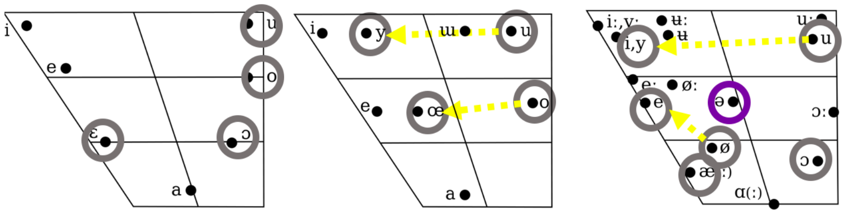


Fig. 5.11: Vowel diagrams of 3 language groups tested. The Norwegian vowel space has at least 9 monophthongs with some of them also having length contrast, such as /ɔ/ vs /ɔ:/. Many dialects of Norwegian have even richer inventories. The central sound that Norwegian has but Italian and Turkish do not, that is /ə/, is denoted inside the purple circle.

5.7.3. Paradigm

The paradigm was the same as the one of the previous experiment (see Section 5.2.3) with a difference in the number of trials. There were in total 256 trials, 64 (16*4) of which were made up of two same vowel sounds. There were 35 pairs that were one edge apart from each other, and they were presented in 140 (35*4) trials. The remaining 52 trials were comparisons between the sounds those that are ‘relatives ’ of each other but more than one edge apart, i.e. comparisons along the /y/-/u/ segment and the /œ/-/o/ segment. 44 (11*4) of those trials were made up of ‘relative’ sounds two edges apart, and 8 (2*4) were three edges apart. For the analyses in the following sections, the data from these 52 trials were not used, and only the 204 trials involving the contrasts of same sounds and contrasts of sounds of one edge apart were considered. For every contrast, say, between the vowel sounds in S1 and S2, half of the presentations had S1 with the voiced and the other half with the voiceless consonant, and S2 with the opposite one. The presentation order of S1 and S2 was swapped so that half of the trials started with S1 and half of them with S2.

We randomized the trials so that two consecutive trials would not have any of the sounds or any of the morphs of those sounds in common, in order to reduce short-term memory effects. As in the previous experiment, every trial started with 1500ms of silence, followed by the presentation of two 400ms long syllables with a 300ms long silence in between them. The participants were asked to respond if they perceived the two sounds as ‘same’ by pressing the space bar in 2000ms, after which a new trial started unless they already made a response before. For the analysis, we accepted the late responses up to 100ms after the 2 seconds response limit as ‘same’.

5.8. Perceptual distances

The number of ‘same’ responses can be represented by a matrix $S(i|j)$, i.e. the fraction of ‘same’ responses when presenting phoneme j and then phoneme i. We can define the perceptual distance between phoneme pairs i and j by first symmetrizing the matrix S, and then by taking the negative log transformation of it:

$$d(i,j) = -ln\left(\frac{S(i|j)S(j|i)}{S(i|i)S(j|j)}\right) \tag{5.1}$$

Note that $d(i,j)$ is a quasi-distance, as it does not necessarily satisfy the required properties, in particular the triangle inequality; but it is sufficient for our purposes.

In Fig. 5.12 we plot the ‘physical distances’ (the Euclidian distances in Bark coordinates) on the plane between phoneme pairs (those that are connected with an edge, here excluding the pair $3-13$ that is an outlier with the largest physical distance) versus the perceptual distances between them. Similar physical distances, i.e. dissimilarity, between the pairs, does not automatically result in similar perceptual distance between them, as the linear fits show. [1] For instance, the pair of the sounds 2 and 3 which has the largest distance in Bark coordinates compared to the other pairs in Fig. 5.12 is not judged to be as the most distant pair, and all three language groups perceive it to be in relative proximity. Indeed, the sounds in the pair $10-11$ that has approximately the same distance as the ones in $2-3$ are perceived to be much different than each other especially by Norwegian and Italian listeners.

The same figure shows us that overall Norwegian speakers are better at discriminating the pairs (13 pairs with $d > 5.33$) than Italian and Turkish speakers (7 pairs each with $d > 5.33$). For a further comparison of perceptual distances of different language groups, we calculated the sum of absolute differences between the perceptual distances for every two mother tongues:

$$SAD = \sum_{p=1}^{35} \frac{\mid (d_{r_p} - d_{l_p}) \mid}{35} \tag{5.2}$$

where d_r and d_l are the perceptual distances of two different groups. In order to determine the maximum bound the difference between two groups can reach, the same measure, $\widetilde{SAD}$, is calculated, but this time from the shuffled distributions of the perceptual distances for every language. The lower the ratio of SAD over $\widetilde{SAD}$, the more similar two distributions are. The calculated ratio is the smallest between Italian and Turkish maps, 0.386, and bigger between Norwegian-Italian and Norwegian-Turkish maps, which are 0.611 and 0.568 respectively.

[1] The same observation is noted in the talk (only abstract available) by [Terbeek and Harshman, 1972]. “In continuation of an experiment described previously [Terbeek and Harshman, 1971], judgments of perceptual similarity and difference for 12 natural vowel sounds were obtained from native speakers of Turkish and Swedish. ..The patterns which emerge suggest that the function relating perceived distances between vowels to their positions along underlying perceptual dimensions is non-Euclidean in two ways. First, the perceptual dimensions do not lie orthogonally to one another, implying that they are related in meaning. Second, they interact nonlinearly, producing a *curved space*, an effect which causes the extraction of an uninterpretable dimension when linear models are used.”

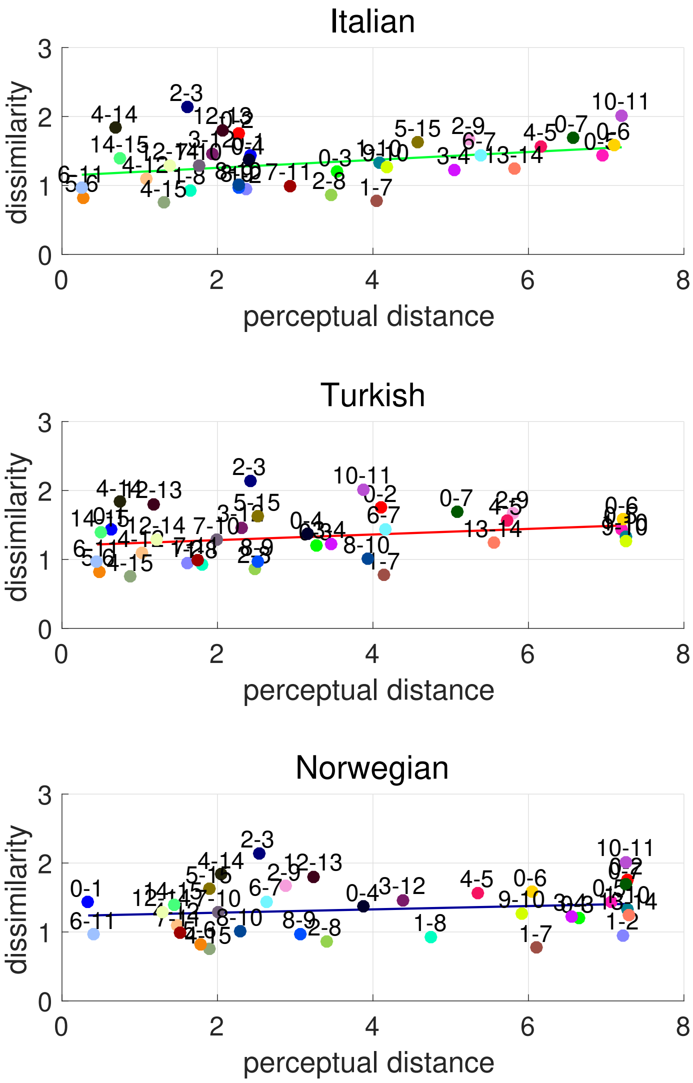


Fig. 5.12: Physical vs. perceptual distances across 3 language groups. Perceptual distances do not follow physical distances in any of the language groups. Different pairs exhibiting close similarities are perceived with varied distances from close to far apart. Norwegian speakers are better at telling different sounds apart, hence they have more of their pairs right with high perceptual distance. The lines are the fits of least squares and the slope of the regressor is flattest for Norwegian speakers and steepest for Italian speakers.

5.9. Self-organizing maps

5.9.1. Algorithm

By making use of an algorithm that 'pushes' the phonemes into a configuration in which their distances are as close as possible to the perceptual ones, we can obtain a re-mapping of the sounds for each language [Kohonen, 1982]. This provides us intuition about language-specific deformation of the vowel plane. The solution found by the algorithm accommodates as well as possible the prescribed distances, but not all distance matrices are exactly reproducible in 2D, especially when the distance matrix does not respect the triangle inequality. The algorithm proceeds as follows.

The initial fit distance between the sounds i and j, that is $f(i,j)$, equals the Euclidean distance between their coordinates, given by the first and second formants F_1 and F_2:

$$f(i,j) = \sqrt{(F_1(i) - F_1(j))^2 + (F_2(i) - F_2(j))^2} \tag{5.3}$$

In order to extract the perceptual distances in the same order of magnitude as the distances in the Bark space, we initially scale them by a scaling factor k:

$$k = \frac{\sum_{i,j \neq i} C(i,j) f(i,j)}{\sum_{i,j \neq i} C(i,j) d(i,j)}. \tag{5.4}$$

where C is the connectivity matrix defined by the 35 edges between the sounds as shown in Fig. 5.10. The scaling factor k of Italian, Turkish and Norwegian is 0.503, 0.516, and 0.402 respectively. If perceptual distances are on average big, we need a smaller scaling coefficient k, and therefore a lower value for k corresponds to a higher ability to discriminate sounds, which is the case for the Norwegian listeners.

At each iteration, a sound i that is different from the previous one is randomly chosen. Its coordinates are then adjusted to minimize a cost function E, which is defined as:

$$E = \sum_{j \neq i} C(i,j) \left[\frac{f(i,j) - d(i,j)}{d(i,j)} \right]^2 \tag{5.5}$$

That is, we do gradient descent to find the local minima, with learning rate α:

$$\Delta F_{1,2}(i) = \alpha \sum_{j \neq i} C(i,j) \left[\frac{(f(i,j) - d(i,j))(F_{1,2}(j) - F_{1,2}(i))}{d(i,j)^2 f(i,j)} \right] \tag{5.6}$$

where we set $\alpha = 0.01$.

5.9.2. Language-dependent maps

The left panel of the Fig. 5.13 shows the physical configuration we start with to obtain language-specific 2D arrangement of the sounds by means of SOM. These 2D arrangements are the "best flattened maps" in the sense that they reflect the perceptual distances as much as possible. We do not here show these maps we obtained, but we assessed the similarity among them by taking one of the them as the reference, and then calculating the deviation of the 35 distance values in another map from the ones in the reference map:

$$RMS = \sqrt{\frac{\sum_{i=1}^{35} (d_{r_i} - d_{l_i})^2}{35}} \tag{5.7}$$

where d_{r_i} is the distance in the reference map between the pair that makes the edge i (i.e. length of the edge i), and d_{l_i} is the same in the map of another language that we compared with the reference map.

The root mean square of the deviations between the Norwegian map and the Turkish map is 0.68, indicating a medium level of similarity, and between Norwegian and Italian maps is 1.03, indicating a larger deviation. The same quantity calculated between Turkish and Italian maps equals to 0.68, suggesting that the deviation between Italian and Norwegian maps are due to few edges with large differences in the perceptual distances, as we will see next. When the *RMS* values are compared to the similarities between perceptual distances of different mother tongue groups, that is given by $\frac{SAD}{\widetilde{SAD}}$ as explained in Section 5.8, we see that the resemblance between the actual perceptual distances is not necessarily preserved in the solutions obtained by the algorithm, some information is lost when the representation is imposed on the 2D plane.

The right panel of Fig. 5.13 is the average map that is created by submitting to the algorithm the aforementioned perceptual distances of the three languages and then by averaging the three maps obtained for each language. We take this average map as the initial setting for each language to prevent the danger of the algorithm amplifying the differences, and then again modify the coordinates to match the perceptual distances of each language. We then can see the differences among the language-specific transformations, as shown in Fig. 5.13. These relative maps maintain the similarities we observed between the first solutions, as we discussed in the previous paragraph.

The final relative maps we obtained are shown in Fig. 5.14. The most noticeable difference between the three maps is the placement of the central sound, which is denoted by 0. In Norwegian, the central sound is kept in the center, slightly pushed down, together with the /œ/ sound, which is attracted towards it. The other language that has the /œ/ category, Turkish, also keeps the central sound close to the /œ/, but slightly separated and upwards. In Italian, where there is no /œ/, the central sound is shifted to the right cluster of sounds, and well separated from /œ/. In the previous experiment, Italian speakers had identified the /œ/ and the central sound, as if for them they were highly similar (see Fig. 5.9). Comparing it with the other nearby sounds, perhaps, made them refine their discrimination in the present experiment.

The right side of the plane is squeezed both in Italian and Turkish maps. In the obtained Italian plane, the sound denoted by 2, is placed closer to its relatives 3 and 13 (/u/), while in the Turkish map, and in particular in the Norwegian map, it is put somewhere in the middle of that segment. The sounds /ɔ/ and /o/ in the right portion are pushed close in the map of Italian speakers, who however have them as distinct categories in their own phoneme inventory. The same is true also for the map of Turkish speakers, whose vowel inventory involves only the /o/ category of two categories, and in fact even closer, but they are more separated in the Norwegian map although Norwegian speakers only possess /ɔ/.

Close to /ɔ/, there is the segment formed by /æ/, /ʌ/, and their morphs, all of which are pushed together to the bottom of the Italian and Turkish maps. The half way morph denoted by 6 is mapped close to /æ/, and away from /ʌ/. Interestingly, while in the previous experiment all the language groups often confused /ʌ/ with /ɔ/, in this experiment, when the participants contrasted the two sound with the other parts of the plane, they found some other sounds to be closer to each one of the /ʌ/ and /ɔ/, hence they are pushed towards opposite clusters, except for Norwegian listeners, on whose map /ʌ/ is kept in between the sound denoted by 6 and /ɔ/.

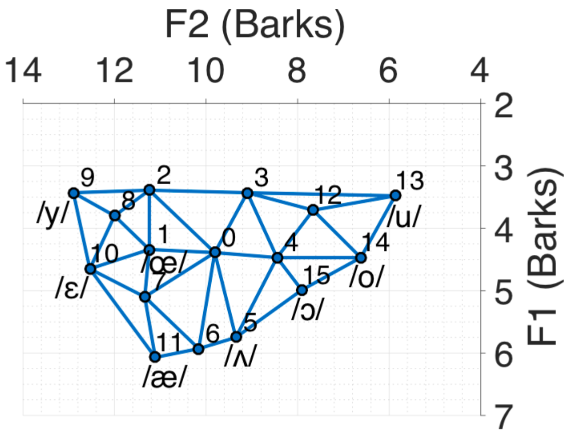


Fig. 5.13: The average maps of 3 mother tongue groups. The left panel shows the initial physical configuration of the sounds. The one on the right panel is the average map that is created by feeding the algorithm with the perceptual distances of the three languages and then by averaging the three maps obtained for each language.

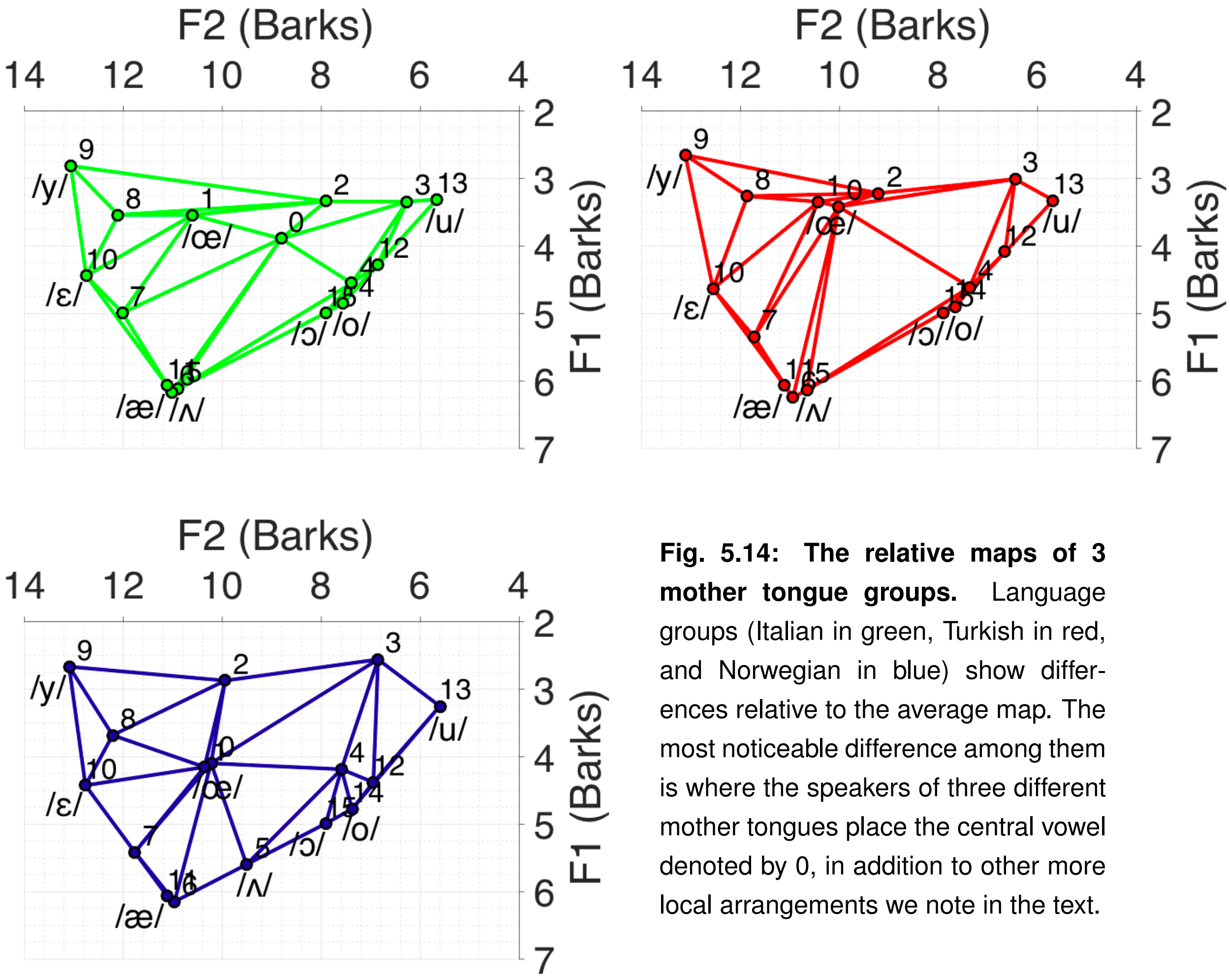


Fig. 5.14: The relative maps of 3 mother tongue groups. Language groups (Italian in green, Turkish in red, and Norwegian in blue) show differences relative to the average map. The most noticeable difference among them is where the speakers of three different mother tongues place the central vowel denoted by 0, in addition to other more local arrangements we note in the text.

Overall, these vowel maps we present here are in contrast with the planes of consonants we presented in Chapter 2.4. The manner and place planes of consonants were in greater agreement across different language groups (with some exceptions) than the ones we observe for vowels, which seem to be language dependent. Then is it possible that speakers, who master two languages, that is bilinguals, maintain two distinct representations of the vowel plane and switch between different vowel maps according to context?

5.10. Multiple maps

So far we have focused on the perceptual and neural representation of vowel plane by non-bilingual speakers, with competing degrees of English competence, or with no/very limited other language experience as in the experiment we presented in Chapter 4. We expect their vowel representation to be mainly shaped by the memory codes of the mother tongue and less influenced by other languages, of which the speaker has no secure command. Here instead we consider the case of native speakers of multiple languages, each of which can be regarded as a single putative mental map. Are bilinguals, who have acquired two languages from their early childhood, in possession of multiple maps [Genesee, 1989, Gonzales and Lotto, 2013, Byers-Heinlein et al., 2017]? Alternatively, are they in possession of one map [Grosjean, 1989] that has a configuration that is unique, or influenced by either each language or by one dominant language?

Current linguistic theories on bilingualism has not settled down the issue yet. Continuing the analogy we make naturally between spatial navigation in 2D arenas and 2D vowel space, we can turn to the findings on the neural representation of multiple spatial contexts to gain an insight of the 'bilingual phenomenon'. Spatial memories of distinct environments are represented by multiple discrete place cell maps [Alme et al., 2014] and which map is active depends on the external sensory cues available at the time [Leutgeb et al., 2005]. When an instantaneous transformation of the spatial context takes place, like switching off the light cues of the enclosure the rat is in, the representation by place cells does not always change all at once, but sometimes goes through a brief period of flickering between the neural maps of the past and the present context in one theta cycle [Jezek et al., 2011]. Is the experience of a bilingual speaker, when she interacts with the speakers of both her mother tongues simultaneously, similar to the one of the rat which is 'teleported' from one environment to another just by a simple cue change?

Hypothesizing that indeed each language produces its distinct metric on the plane, as the results we presented in Section 5.6 and also in Section 5.1 suggest, we can further ask whether native bilingual listeners, who can effortlessly parse phonemes drawn from incompatible repertoires, use separate metric representations for each language, or some kind of mixed, refined metric on the same representation, that allows them to decode any relevant vowel contrast in either language. Can there be a neural mechanism similar to the one observed in the hippocampal CA3 network [Jezek et al., 2011] underlying either a rapid perceptual switch in response to sudden changes in the cues that define the language context? In the domain of production, it was observed that the articulatory system is flexible between the two languages of bilinguals and the switch from one language to another involves a total change at the phonetic level [Grosjean and Miller, 1994].

Here we present the results from a pilot experiment we conducted not with bilinguals, but with native Norwegian speakers who are fluent in English. Can speakers who have excellent command of multiple languages develop distinct perceptual maps? Is that what makes them to be skilled in foreign languages? If they do possess different maps, then understanding the neural computations endowing these representations become even more relevant as the world we live in becomes more global and we are required to interact in different languages more than any other time in history. In a future work, we would like to include true bilingual populations and compare their representations with excellent second language speakers.

5.11. Methods and materials

5.11.1. Stimuli

The same stimuli as in Section 5.7.1 were used. For the context exposure paradigm explained below, a native British male speaker and a native Norwegian female speaker recorded three short stories in their respective mother tongues. The English stories were 'Ex Oblivione' by H.P. Lovecraft, 'Coming of the King' by Nikolai Tolstoy , and 'A Blunder' by Anton Chekhov. The Norwegian stores were 'God dag mann! Økseskaft!' ('Good day, fellow! Axe handle'), 'Mannen som skulle stelle hjemme' ('The husband who was to mind the house'), 'Kjerringa mot strømmen' ('The wife above the waterfall') all written by Peter Christen Asbjørnsen and Jørgen Moe. The stories were cut at a meaningful point to have a length of four, two and four minutes in order.

5.11.2. Participants

We tested 6 university students who are native Norwegian speakers, and had very high self-rated English comprehension. Based on their academic backgrounds, we accept it to be an accurate report.

5.11.3. Paradigm

The paradigm is the same as the one explained in Section 5.7.3 with a difference that the participants were exposed to a short story session before they started a test session. The first story session was the longest in length, i.e. 4minutes, as it was the initial exposure. The following story session was comprised of the second story ('Coming of the King'/'Mannen som skulle stelle hjemme') and lasted 2 minutes long. The last short story ('A Blunder'/'Kjerringa mot strømmen') was divided into two in the middle, thus the participants listened to the first half in the third session, and the other half in the last session, each were 2 minutes long. In order to make sure the participants were actively listening to the stories, they were asked to answer yes-no questions at the end of each story session. The number of questions were proportional to the length of the story they heard and there were 10 questions in total. 256 trials (64*4) were distributed equally among 4 test sessions, and presented in the same fashion as in Section 5.7.3.

5.12. One map for all

The non-bilingual participants had a high percentage of answering the quiz questions correctly for English stories, 95% on average, but much lower for Norwegian stories, 46.67% on average. Perhaps, they were not very careful and attentive to the stories in their mother tongue, but much more so to the stories in a foreign language, as they wanted to show that their English comprehension level matches their self-report of their fluency in English. Therefore, we assume they were attentive to the contextual setting, at least to the one that was not the their mother tongue. In fact as their mother tongue is Norwegian, and they are not bilingual, their default map should reflect the one shaped by Norwegian. We will see next if the difference in the settings was enough to trigger changes in the behavior.

We made use of the same algorithm explained in Section 5.9 (with learning rate α set to 0.008) to produce the two maps of the participants after they were exposed to each language, as shown in Fig. 5.15. The two maps show very similar arrangements with minimal differences, which are mostly due to the fact that small differences are magnified as there were only 6 participants.

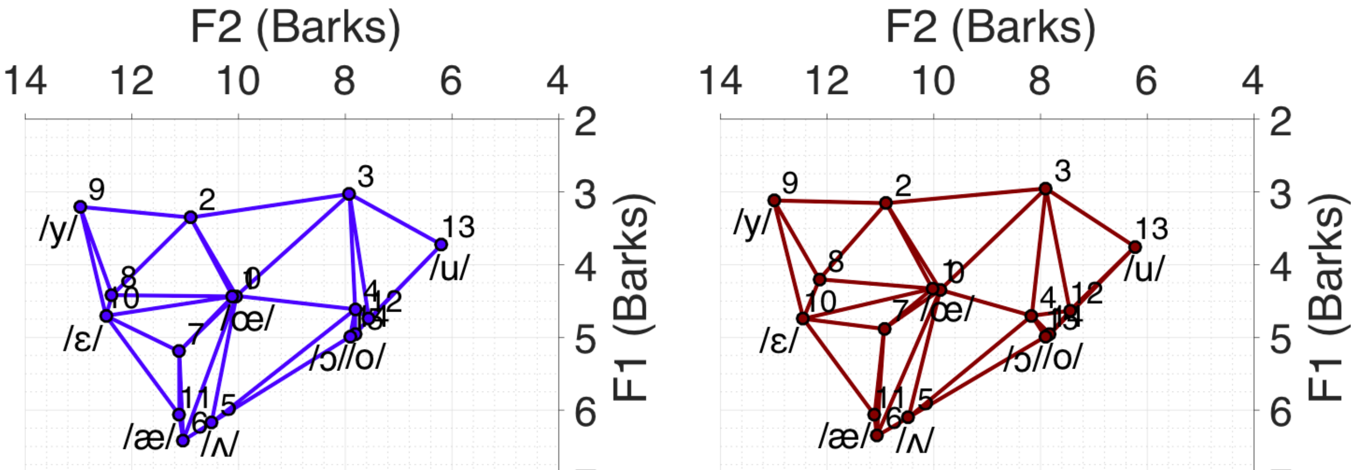


Fig. 5.15: The context-independent maps of non-bilingual Norwegian speakers. The deformation of the perceptual map of non-bilingual native Norwegian speakers after listening to the stories in English (the map on the right panel) look similar to their default map (the one on the left panel) which we obtain after they are exposed to Norwegian stories. The central sound denoted by 0 is kept in the center in both maps, as it was by Norwegian speakers in the previous experiment (see Fig. 5.14).

We quantified the similarity of each one of the two maps to the previously obtained Norwegian map and the similarity among the two maps here, the same way we did for Turkish, Italian and Norwegian maps in Section 5.9.2. The non-bilingual native Norwegian speakers, who are fluent in

English, do not alter their representations as the calculated small deviation from each other, that is 0.17, indicates. The deviation (RMS) of the map after Norwegian exposure (the left panel of Fig. 5.15) from the Norwegian map produced without the context paradigm is 0.45, and the same for the map after English exposure is 0.48, both suggesting a higher level of similarity than the previously reported 0.68 between Turkish and Norwegian.

The two maps in Fig. 5.15 are in agreement on where they place the central vowel denoted by 0 with the Norwegian map we obtained previously (see Fig. 5.14). However, some local structures differ; for example, the pentagon formed by the sounds denoted by 10,1,0,6, and 11 at its vertices. In the previous experiment it got the shape of a triangle, where the sounds denoted by 10, 11 (with the sound 6), 1 (with the sound 0) occupy the vertices. Without the context-switch paradigm, the sound 7 lays on the edge connecting the two vertices 10 and 11 (with 6), whereas with the context-switch paradigm it lies at the center of the triangle formed by the same vertices. A further difference is that the distance between the sounds /ʌ/ and /ɔ/ is slightly increased and the two phonemes are slightly dragged towards the opposite sides. This might be caused of the dialect differences between the participants of two experiments; the two groups of Norwegian speakers were tested in two different cities: Tromsø for the former experiment and Trondheim for the latter. We need to test more participants to understand if it reflects a perceptual difference due to the difference in the dialect or not.

Seeing that, for non-bilingual speakers with a high-level of English fluency, the representation remains stable after experiencing different contexts (in the situation studied here, at least) shows us, once again, how solid one's perceptual map is once it is developed, which happens very early in life during what is called as *critical phonetic period* [Houston et al., 2007, Kuhl et al., 2003]. Indeed, the similarity between the maps of non-bilingual speakers (see Fig. 5.15) makes the perceptual maps of bilinguals who possess the phonetic contrasts of two languages, even more intriguing.

6

General Discussion

Traditionally, phoneme representation in the brain, like other components of language processing, has been studied with dendrograms of hierarchically clustered discrete items [Bouchard et al., 2013]. Accordingly, many aspects of language competence have been associated with setting the corresponding discrete or even binary values for an ensemble of language parameters describing the internal structure of that language. In the domain of vowel production and perception, the situation is much less clear. Phonologists have described the standard vowels used in a given language in terms of binary parameters. For example, as we discussed in Chapter 3, in Turkish the 3 binary parameters [open/closed], [front/back], [unrounded/rounded] suffice to characterize the 8 standard vowels, which sit at the 8 vertices of the resulting cube. In other languages, however, the parametric description is less convincing, and many vertices remain unoccupied.

On the other hand, phonetically vowels admit an accurate description as points in the two dimensional plane spanned by the first and second formant, which correlate outstandingly well with the phonological (production) parameters [open/closed] and [front/back] but extend them to a continuous range. The same is not true for consonants; despite the fact that they are also characterized by various binary values [voiced/voiceless],[plosive/fricative],[aspirated/not aspirated],[labial/velar] (except for the non-binary voice-onset timing for plosives in some languages [Laufer, 1998]), these values do not reflect a continuous physical organization in their production, thus they are too complex to be represented by a simple two dimensional grid.

We took advantage of the continuous 2D Euclidean space of vowels and investigated its mental representation. Previous studies that investigated vowel perception made use of synthesized vowel continua that morphed in equal steps, and they focused on vowel categorization with paradigms like AxB two forced-choice task [Pisoni, 1973,Eerola and Aaltonen, 2012,Silvia and Rothe-Neves, 2016]. Instead, we used a different approach and we were interested in obtaining the perceptual metric of the space; that means not directly in the categorization itself, but whether that metric was faithful to the one used in the production.

The years we have studied phoneme perception coincided with the period that the Nobel Prize in Physiology and Medicine was awarded for the discovery of spatial maps in the brain, a concept which attracted the increased interest of a larger community in neuroscience, which investigates different brain functions. The circuitry for spatial navigation contains grid cells with firing fields that tile the entire environment in a periodic hexagonal pattern [Fhyn et al., 2008]. Subsequently, grid-like representations have been reported in humans in navigation tasks that took place in virtual environments [Doeller et al., 2010], and also even in an abstract 2D space [Constaninescu et al., 2016]. Furthermore, grid-like activity was not limited to the entorhinal cortex but observed in variety of regions including the ventra-medial prefrontal cortex, suggesting a widespread grid code. We investigated whether similar computational principles extend to the processing of another 2D manifold like the vowel plane, representation of which take place in the sensory auditory cortex to which the entorhinal cortex feeds back [Chen et al., 2013].

Because the presence of borders deforms the grid expression [Stensola et al., 2015], we reasoned that if the perceptual metric corresponds to the one of production, then the central region of the Italian vowel space, which has no vowels inside, is the most promising portion to reflect a Euclidean metric. We trained the participants over long training-sessions so that we could achieve an equally good discrimination, approximately, of all the trajectories, which was necessary for a grid-like representation to emerge. However, in contrary to our expectation, the underlying neural representation did not express hexagonal symmetry We found, instead, a suggestion of a sensitivity of the neural signal mainly to changes in opposite directions in the resonance frequencies along the plane.

Seeing that evoked potentials showed the greatest variation for the trajectories that are in between the positions of Italian vowels around the outer contour of the wheel, we wondered how do the vowel categories/boundaries influence the metric perception of the vowel plane? A similar paradigm we used in the 'wheel experiment' in Chapter 4 could be adapted to study the language-specific geometry of the plane systematically, and how they relate to the different vowel categories existing in a given vowel repertoire and to their usage frequencies, if there is such a transparent and simple relationship. One can move the wheel around the entire vowel space and position it at different centers, change the number of trajectories, and even alter its size. However, when the 'wheel' is used to explore the nature of the representation, as opposed to flattening it as much as possible as we did in our experiment, no intensive training would be required. This is left for future work.

To study the effect of putative vowel attractors on the metric of vowel space, we employed another paradigm, in which we connected distant sounds through small steps of morphs which do not correspond to vowel category on their own. We mapped the complex relations among the sounds in the high-dimensional space into simple geometric relations by a dimension reduction technique [Kohonen, 1982]. Previously, using the same algorithm, the brain based representation of distant vowels (/i/ /u/ /o/ in one study, and /i/ /u/ /a/ in the other) was analysed [Formisano et al., 2008, Bonte et al., 2014] and a metric relationship was reported between the distances of the cortical representation of the vowels and their physical distance in the space. However, perceptual distances of shorter physical distances tell more about the curvature of the vowel plane. Geometrical relations are distorted as there is a perceptual deformation that attracts sound around a prototype of the standard category towards itself [Kuhl, 1991, Oudeyer, 2006, Lui et al., 2005]. A neural network model explained this effect as a result of the non-uniform distribution of firing preferences of auditory neurons caused by the exposure to a particular distribution of sounds, i.e., to the language-specific phoneme inventory [Guenther and Gjaja, 1996].

Accordingly, our results suggest that, presumably due to the influence of language-specific vowel categories, perception of distances in the vowel space do not follow metric relations. Rather, perception is dynamically changing according to which segment and which language you are 'walking' in. For example, in the Turkish perceptual map, a bigger jump is required to reach from /ɛ/ to

/œ/ compared to the perceptual step size that has to be taken in the Italian map. This elasticity of the plane from one language to another is in contrast to the consonant ‘planes’ we presented in Chapter 2, which remained mainly stable across language groups, hinting at different cross-linguistic structures of consonant and vowel representation. It is left for future studies to study the neural representation of the deformations of vowel space.

Within the assumption that transforming speech signals to phonemes is a critical step in comprehending speech, our research aims to understand perceptual and neural phoneme representations, in order to afford insights into the cortical space which represents all possible vocalizations in one’s own native language, and into the trajectories which neural activity can follow in this space as one is continuously exposed to highly variable sound patterns.

Bibliography

[Alme et al., 2014] Alme, C. B., Miao, C., Jezek, K., Treves, A., Moser, E. I., and Moser, M.-B. (2014). Place cells in the hippocampus: Eleven maps for eleven rooms. *PNAS*, 111(52):18428–18435.

[Amit, 1989] Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks*. Cambridge University Press, New York.

[Arsenault and Buchsbaum, 2015] Arsenault, J. S. and Buchsbaum, B. R. (2015). Distributed neural representations of phonological features during speech perception. *The Journal of Neuroscience*, 35(2):634–642.

[Balas, 2009] Balas, A. (2009). Why can poles perceive Sprite but not Coca-Cola? a natural phonological account. In Boersma, P. and Hamann, S., editors, *Phonology in perception*, chapter 2,, pages 25–54. Mouton de Gruyter, Berlin.

[Bassetti and Atkinson, 2015] Bassetti, B. and Atkinson, N. (2015). Effects of orthographic forms on pronunciation in experience instructed second language learners. *Applied Psycholinguistics*, 36(01):67–91.

[Blumeyer, 2012] Blumeyer, D. (2012). English phoneme frequencies.

[Boatman et al., 1997] Boatman, D., Hall, C., Goldstein, M. H., Lesser, R., and Gordon, B. (1997). Neuroperceptual differences in consonant and vowel discrimination: as revealed by direct cortical electrical interference. *Cortex*, 33:83–98.

[Boersma and Weenik, 2018] Boersma, P. and Weenik, D. (2018). Praat: doing phonetics by computer.

[Bonte et al., 2014] Bonte, M., Hausfeld, L., Scharke, W., Valente, G., and Formisano, E. (2014). Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *The Journal of Neuroscience*, 34(13):4548–4557.

[Bouchard et al., 2013] Bouchard, K. E., Mesgarani, N., Johnson, K., and Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495:327–332.

[Boyce, 1990] Boyce, S. E. (1990). Coarticulatory organization for lip rounding in turkish and english. *The Journal of the Acoustical Society of America*, 88(06):2585–2595.

[Brannen, 2011] Brannen, K. J. (2011). *The perception and production of interdental fricatives in second language acquisition*. PhD thesis, McGill University.

[Bruneau et al., 1997] Bruneau, N., Roux, S., Guerin, P., Barthelemy, C., and Lelord, G. (1997). Temporal prominence of auditory evoked potentials (n1 wave) in 4-8-year-old children. *Psychophysiology*, 34(1):32–38.

[Buck, 1970] Buck, J. H. (1970). The influence of sanskrit on the japanese sound system. Presented at the Southeastern Conference on Linguistics.

[Bullmore et al., 1999] Bullmore, E. T., Suckling, J., and Overmeyer, S. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural mr images of the brain. *IEEE transactions on medical imaging*, 18(1):32–42.

[Byers-Heinlein et al., 2017] Byers-Heinlein, K., Morin-Lessard, E., and Lew-Williams, C. (2017). Bilingual infants control their languages as they listen. *PNAS*, 114(34):9032–9037.

[Caramazza et al., 2000] Caramazza, A., Chalant, D., Capassio, R., and Miceli, G. (2000). Separable processing of consonants and vowels. *Nature*, 403(27):428–430.

[Chaumon et al., 2015] Chaumon, M., Bishop, D. V. M., and Busch, N. A. (2015). A practical guide to the selection of independent components of the eectroencephalogram for artifact correction. *Journal of neuroscience methods*, 250:47–63.

[Chen et al., 2013] Chen, X., Guo, Y., Feng, J., Liao, Z., Xinjian, L., Wang, H., Li, X., and He, J. (2013). Encoding and retrieval of artifical visuoauditory memory traces in the auditory cortex requires the entorhinal cortex. *Journal of Neuroscience*, 33(24):9963–9974.

[Ciaramelli et al., 2006] Ciaramelli, E., Lauro-Grotto, R., and Treves, A. (2006). Dissociating episodic from semantic access mode by mutual information measures: Evidence from again and alzheimer's disease. *Journal of Physiology Paris*, 100(81):142–153.

[CLARIN-NL TDS Curator, 2018] CLARIN-NL TDS Curator, . (2018). The ucla phonological segment inventory database.

[Constaninescu et al., 2016] Constaninescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.

[Cooke et al., 2010] Cooke, M., Lecumberri, M. L. G., Scharenborg, O., and van Dommelen, W. A. (2010). Language-independent processing in speech perception: identification of english intervocalic consonants by speakers of eight european languages. *Speech communication*, 52:954–967.

[Cutler et al., 2004] Cutler, A., Weber, A., Smits, R., and Cooper, N. (2004). Patterns of english phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, 116(6):3668–3678.

[de Boer, 1999] de Boer, B. (1999). Evolution and self-organisation in vowel systems. *Evolution of Communication*, 3(1):79–102.

[de Saussure, 1966] de Saussure, F. (1966). *Course in General Linguistics*. McGraw-Hill, New York.

[Dehaene−Lambert, 1997] Dehaene−Lambert, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *Neuroreport*, 8(4):919–924.

[Delorme and Makeig, 2004] Delorme, A. and Makeig, S. (2004). Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21.

[Derdikman et al., 2009] Derdikman, D., Whitlock, J. R., Tsao, A., Fhyn, M., May-Britt, M., and Moser, E. I. (2009). Fragmentation of grid cell maps in multicompartment environment. *Nature Neuroscience*, 12(10):1325–1332.

[DiCanio et al., 2015] DiCanio, C., Nam, H., Amith, J. D., Castillo Garcia, R., and Whalen, D. H. (2015). Vowel variability in elicited versus spontaneous speech: Evidence from mixtec. *Journal of Phonetics*, 48:45–59.

[Doeller et al., 2010] Doeller, C. F., Barry, C., and Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, 463:657–661.

[Eerola and Aaltonen, 2012] Eerola, O. and Aaltonen, O. (2012). The effect of duration on vowel categorization and perception prototypes in a quantity language. *Journal of Phonetics*, 40(2):315–328.

[Eimas and Corbit, 1973] Eimas, P. D. and Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 4:99–109.

[Ferro et al., 1978] Ferro, F. E., Magno-Caldognetto, E., Vagges, K., and Lavagnoli, C. (1978). Some acoustic characteristics of italian vowels. *Journal of Italian Linguistics*, 3(1):87–95.

[Fhyn et al., 2007] Fhyn, M., Hafting, T., Treves, A., May-Britt, M., and Moser, E. I. (2007). Hippocampal remapping and grid realignment in entorhinal cortex. *Nature*, 446(7132):190–194.

[Fhyn et al., 2008] Fhyn, M., Hafting, T., Witter, M. P., Moser, E. I., and Moser, M.-B. (2008). Grid cells in mice. *Hippocampus*, 18(12):1230–1238.

[Files et al., 2015] Files, B. T., Tjan, B. S., Jiang, J., and Bernstein, L. E. (2015). Success and failure of new speech category learning in adulthood: Consequences of learned hebian attractors in topograhic maps. *frontiers in Psychology*, 13(6):878.

[Flege, 1995] Flege, J. E. (1995). Second languages speech learning: Theory, findings, and problems. In Di Sciullo, A. M., editor, *Speech perception and linguistic experience: issues in cross-language research*, chapter 8, pages 233–277. York Press, Timonium.

[Formisano et al., 2008] Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" is saying "what"? brain-based decoding of human voice and speech. *Science*, 322(5903):970–973.

[Fox, 1982] Fox, R. A. (1982). Individual variation in the perception of vowels: implications for a perception-production link. *Phonetica*, 39:1–22.

[Fox, 1983] Fox, R. A. (1983). Perceptual structure of monophthongs and diphthongs in english. *Language and Speech*, 26(1):21–60.

[Fry, 2004] Fry, E. (2004). Phonics: A large phoneme-grapheme frequency count revised. *Journal of Literacy Research*, 36(1):85–98.

[Gay, 1968] Gay, T. (1968). Effect of speaking rate on diphthong formant movements. *The Journal of the Acoustical Society of American*, 44(6):1570–1573.

[Gay, 1970] Gay, T. (1970). A perceptual study of american english diphthongs. *Language and Speech*, 13(1):65–88.

[Genesee, 1989] Genesee, F. (1989). Early bilingual development: one language or two? *Journal of Child Language*, 16(1):161–179.

[Gonzales and Lotto, 2013] Gonzales, K. and Lotto, A. J. (2013). A bafri, un pafri. bilinguals' pesudoword identification support language-specific phonetic systems. *Pyschological Science*, 24(11):2135–2142.

[Goslin et al., 2014] Goslin, J., Galluzzi, C., and Romani, C. (2014). Phonitalia: a phonological lexicon for italian. *Behavior Research Methods*, 46(3):872–886.

[Grimaldi et al., 2016] Grimaldi, M., Di Russo, F., and Sigona, F. (2016). Electroencephalographic evidence of vowels computation and representation in human auditory cortex. In Di Sciullo, A. M., editor, *Biolinguistic investigation on the language faculty*, chapter Language faculty, pages 79–101. John Benjamins Publishing, Amsterdam.

[Grosjean, 1989] Grosjean, F. (1989). Neurolinguists, beware! the bilingual is not two monolinguals in one person. *Brain and Languages*, 36:3–15.

[Grosjean and Miller, 1994] Grosjean, F. and Miller, J. L. (1994). Going in and out of languages: An example of bilingual flexibility. *Psychological Science*, 5(4):201–206.

[Guenther and Gjaja, 1996] Guenther, F. H. and Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *The Journal of the Acoustical Society of America*, 100(2):1111–1121.

[Guirao and Garcia Jurado, 1990] Guirao, M. and Garcia Jurado, M. A. (1990). Frequency of occurrence of phonemes in american spanish. *Revue quëbëcoise de linguistique*, 19(2):135–149.

[Hafting et al., 2005] Hafting, T., Fhynn, M., Molden, S., Moser, M.-B., and Moser, E. I. (2005). Microstructure of a spatial map in the enthorhinal cortex. *Nature*, 436(7052):69–89.

[Hayden, 1950] Hayden, R. (1950). The relative frequency of phonemes in general american english. *WORD*, 6(3):217–223.

[Houston et al., 2007] Houston, D. M., Horn, D. L., Qi, R., Ting, J. Y., and Gao, S. (2007). Assessing speech discrimination in individual infants. *Infancy*, 12:119–145.

[Huang and Johnson, 2010] Huang, T. and Johnson, K. (2010). Language specificity in speech perception: Perception of mandarin tones by native and nonnative listeners. *Phonetica*, 67:243–267.

[Hubel and Wiesel, 1959] Hubel, D. H. and Wiesel, T. (1959). Receptive fields of single neurons in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591.

[IPA, 2018] IPA (2018). The international phonetics alphabet and the ipa chart.

[Iverson et al., 2003] Iverson, P., Kuhl, P., Akahane-Yamade, R., Diesch, E., Tohkure, Y., Kettermann, A., and Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, 87:47–57.

[Jezek et al., 2011] Jezek, K., Henriksen, E. J., Treves, A., Moser, E. I., and Moser, M.-B. (2011). Theta-paced flickering between place-cell maps in the hippocampus. *Nature*, 478:246–249.

[Kaun, 2004] Kaun, A. R. (2004). The typology of rounding harmony. In Hayes, B., Kirchner, R., and Steriade, D., editors, *Phoneticall-Based Phonology*, chapter 4, pages 87–116. Cambridge University Press, Cambridge.

[Kaun, 2010] Kaun, A. R. (2010). Vertical convergence of linguistic varieties in a language space. In Ryneland, U., editor, *Language and Space: Theories and Methods*, chapter 3, pages 259–274. De Gruyter, Germany.

[Kawahara et al., 1999] Kawahara, H., Masuda-Katsuse, I., and de Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time-freqeuncy smoothing and an instantaneuous-frequency based f0 extraction: possible role. *Speech Communication*, 27(3-4):187–207.

[Kendall, 1989] Kendall, D. G. (1989). A survey of the statistical theory of shape. *Statistical Science*, 4(2):87–99.

[Khalighinejad et al., 2017] Khalighinejad, B., da Silva, G. C., and Mesgarani, N. (2017). Dynamic encoding of acoustic features in neural responses to continuous speech. *Journal of Neuroscience*, 37:2176–2185.

[Killian et al., 2012] Killian, N. J., Jutras, M. L., and Buffalo, E. A. (2012). A map of visual space in the primate entorhinal cortex. *Nature*, 491:761–764.

[Klatt, 2013] Klatt, D. H. (2013). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, 67(3):971–995.

[Kohonen, 1982] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

[Kriegeskorte and Storrs, 2016] Kriegeskorte, N. and Storrs, K. R. (2016). Grid cells for conceptual spaces? *Neuron*, 92:280–284.

[Kristoffersen, 2000] Kristoffersen, G. (2000). *The phonology of Norwegian*. Oxford University Press.

[Kropff et al., 2015] Kropff, E., Carmichael, J. E., May-Britt, M., and Moser, E. I. (2015). Speed cells in the medial entorhinal cortex. *Nature*, 523:419–424.

[Kropff and Treves, 2008] Kropff, E. and Treves, A. (2008). The emergence of grid cells: intelligent design or just adaptation? *Hippocampus*, 18:1256–1269.

[Kuhl, 1991] Kuhl, P. 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.

[Kuhl, 2000] Kuhl, P. K. (2000). A new view of language acquisition. *PNAS*, 22:11850–11857.

[Kuhl et al., 2003] Kuhl, P. K., Tsao, F.-M., and Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *PNAS*, 100:9096–9101.

[Laufer, 1998] Laufer, A. (1998). Voicing in contemporary hebrew in comparison with other languages. *Hebrew Studies*, 39:143–179.

[Lehiste and Meltzer, 1973] Lehiste, I. and Meltzer, D. (1973). Vowel and speaker identification in natural synthetic speech. *Language and Speech*, 16:356–364.

[Leutgeb et al., 2005] Leutgeb, J. K., Leutgeb, S., Treves, A., Meyer, R., Barnes, C. A., McNaughton, B. L., Moser, M.-B., and Moser, E. I. (2005). Progressive transformation of hippocampal neuronal representations in "morphed" environments. *Neuron*, 48(2):345–358.

[Liberman et al., 1957] Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boudaries. *Journal of Experiment Psychology*, 54(5):358–368.

[Lisker, 1989] Lisker, L. (1989). On the articulatory interpretation of vowel quality: The dimension of rounding. *Journal of the International Phonetic Association*, 19(1):24–30.

[Lorenzo et al., 2018] Lorenzo, P., Cocco, S., and Monasson, R. (2018). Integration and multiplexing of positional and contextual information by the hippocampal network. *Plos Computational Biology*, 14(8):1–23.

[Lui et al., 2005] Lui, H.-M., Tsao, F.-M., and Kuhl, P. K. (2005). The effect of reduced vowel working space on speech intelligibility in mandarin-speaking young adults with cerebral palsy. *The. Journal of Acoustical Society of America*, 117(6):3879–3889.

[Macpherson, 1975] Macpherson, I. R. (1975). Plosive, fricative and affricate consonants in greater detail. In *Spanish Phonology: Descriptive and Historical*, chapter 8, pages 87–116. Manchester University Press, Manchester.

[Maidenbaum et al., 2018] Maidenbaum, S., Miller, J., Stein, J. M., and Jacobs, J. (2018). Grid-like hexadirectional modulation of human entorhinal theta oscillations. *PNAS*, 115(42):10798–10803.

[Maris and Oostenveld, 2007] Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of eeg and meg data. *Journal of Neuroscience Methods*, 164(1):177–190.

[Marozzi and Jeffery, 2012] Marozzi, E. I. and Jeffery, K. J. (2012). Place, space and memory cells. *Current Biology*, 22(22):939–942.

[Masapollo et al., 2017] Masapollo, M., Polka, L., and Molnar, M. (2017). Directional asymmetries reveal a universal bias in adult vowel perception. *The Journal of the Acoustical Society of America*, 141(4):2857–2689.

[McGurk and MacDonald, 1976] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264:746–748.

[McMurray and Spivey, 2000] McMurray, B. and Spivey, M. (2000). The categorical perception of consonants: the interaction of learning and processing. *Proceedings of the Chicago Linguistics Society*, 34(2):205–220.

[Miller and Nicely, 1954] Miller, G. A. and Nicely, P. E. (1954). An analysis of perceptual confusions among some english consonants. *The journal of the acoustical society of America*, 27(2):338–352.

[Moser et al., 2008] Moser, E. I., Kropff, E., and Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review Neuroscience*, 31:69–89.

[Moser and Moser, 2008] Moser, E. I. and Moser, M.-B. (2008). A metric for space. *Hippocampus*, 18(12):1142–1156.

[Muller et al., 1987] Muller, R. U., Kubie, J. L., and Ranck, J. B. (1987). The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *Journal of Neuroscience*, 7(7):1935–1950.

[Näätänen, 2001] Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (mmn) and its magnetic equivalent (mmnm). *Psychophysiology*, 38(1):1–21.

[Näätänen et al., 1999] Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainia, M., Alku, P., Ilmoniemi, R. J., Luuk, A., Allik, J., Sikkonen, J., and Alho, K. (1999). Language−specific phoneme representation revealed by electric and magnetic brain responses. *Nature*, 385:432–434.

[Obleser et al., 2003] Obleser, J., Elbert, T., Lahiri, A., and Eulitz, C. (2003). Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. *Cognitive Brain Research*, 15:207–213.

[Obleser et al., 2004] Obleser, J., Lahiri, A., and Eulitz, C. (2004). Magnetic brain response mirrors extraction of phonological features from spoken vowels. *Journal of Cognitive Neuroscience*, 16:31–39.

[Oflazer, 2016] Oflazer, K. (2016). Milliyet corpus.

[O'Keefe and Dostrovsky, 1971] O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map. preliminary evidence from unit activity in the freely moving rat. *Brain Research*, 34(1):171–175.

[Oostenveld et al., 2011] Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). Fieldtrip: Open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011(156869).

[Oudeyer, 2006] Oudeyer, P.-Y. (2006). *Self-organization in the evolution of speech*. Oxford University Press, New York.

[Palmer, 2011] Palmer, S. B. (2011). English phoneme frequencies.

[Peterson and Barney, 1952] Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24:175–184.

[Phatak and Allen, 2007] Phatak, S. A. and Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise. *The journal of the acoustical society of America*, 121(4):2312–2326.

[Picton, 1974] Picton, T. W. (1974). Human auditory evoked potentials. i: Evaluation of components. *Electroencephalography and Clinical Neurophysiology*, 36(1):179–190.

[Pisoni and Tash, 1974] Pisoni, D. and Tash, J. (1974). Reaction times to comparisons with and across phonetic categories. *Perception and Psychophysics*, 15(2):285–290.

[Pisoni, 1973] Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, 13(2):253–260.

[Polka and Bohn, 2003] Polka, L. and Bohn, O.-S. (2003). Asymmetries in vowel perception. *Speech Communication*, 41(1):221–231.

[Redford and Diehl, 2004] Redford, M. A. and Diehl, R. L. (2004). The relative perceptual distinctivemess of inital and final consonants in cvc syllables. *The Journal of the Acoustical Society of America*, 116(6):3668–3678.

[Roach, 2004] Roach, P. (2004). British english: Received pronunciation. *Journal of the International Phonetic Association*, 34(3):239–245.

[Roberts et al., 2000] Roberts, T. P. L., Ferrari, P., Stufflebeam, M., and Poeppel, D. (2000). Latency of the auditroy evoked neuromagnetic fields components: stimulus dependence and insights toward perception. *Journal of Clinical Neurophysiology*, 17(2):114–129.

[Russell, 2009] Russell, K. (2009). Vowel formants.

[Saenz-Badillos, 2002] Saenz-Badillos, A. (2002). *A history of the Hebrew Language*. Cambridge University Pressl, Cambridge.

[Schwartz et al., 1997] Schwartz, J.-L., Bo , Louis-Jean, V., and Abry, C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, 25:255–286.

[Silber-Varod, 2011] Silber-Varod, V. (2011). Hebrew phoneme frequencies. Tel-Aviv Academic College for Engineering, Afeka Center for Language processing.

[Silvia and Rothe-Neves, 2016] Silvia, D. M. R. and Rothe-Neves, R. (2016). Perception of height and categorization of brazilian portuguese front vowels. *Journal of Phonetics*, 32(2):355–373.

[Solstad et al., 2008] Solstad, T., Boccara, C. N., Kropff, E., May-Britt, M., and Moser, E. I. (2008). Representation of geometric borders in the entorhinal cortex. *Science*, 322(5909):1865–1868.

[Sproge, 2003] Sproge, D. (2003). A comparison of phonetic characteristics of diphthongs in latvian and lithuanian.

[Staudigl et al., 2018] Staudigl, T., Leszczynki, M., Jacobs, J., Sheth, S. A., E, S. C., Jensen, O., and Doeller, C. F. (2018). Hexadirectional modulation of high-frequency electrophysiological activity in the human anterior medial temporal lobe maps visual space. *Current Biology*, 28:1–5.

[Stella et al., 2013] Stella, F., Cerasti, E., and Treves, A. (2013). Unveiling the metric structure of internal representations of space. *Frontiers in Neural Circuits*, 7(81).

[Stensola et al., 2012] Stensola, H., Stensola, T., Solstad, T., Frland, K., May-Britt, M., and Moser, E. I. (2012). The entorhinal grid map is discretized. *Nature*, 492:72–78.

[Stensola et al., 2015] Stensola, T., Stensola, H., Moser, M.-B., and Moser, E. I. (2015). Shearing-induced asymmetry in entorhinal grid cells. *Nature*, 518:207–212.

[Taube et al., 1990] Taube, J. S., Muller, Robert, U., and Ranck, J. B. J. (1990). Head-direction cells recorded from the postsubiculum in freely moving rats. i. description and quantitative analysis. *Journal of Neuroscience*, 10(2):420–435.

[ten Pever and Sack, 2015] ten Pever, S. and Sack, A. T. (2015). Oscillatory phase shapes syllable perception. *PNAS*, 112(52):15833–15837.

[Terbeek and Harshman, 1971] Terbeek, D. and Harshman, R. (1971). Cross-language differences in the perception of natural vowel sounds. *The Journal of the Acoustical Society of America*, 50(1A):147–147.

[Terbeek and Harshman, 1972] Terbeek, D. and Harshman, R. (1972). Is vowel perception non-eucledian? *The Journal of the Acoustical Society of America*, 51(1A):81–81.

[Thompson, 1942] Thompson, D. W. (1942). *On Growth and Form*. Cambridge University Press.

[Tolman, 1948] Tolman, E. C. (1948). Cognitive maps in rats and men. *The Psychological Review*, 55(4):189–208.

[Torgerson, 1952] Torgerson, W. S. (1952). Multi-dimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419.

[Treves, 1997] Treves, A. (1997). On the perceptual structure of face space. *BioSystems*, 40:189–196.

[Tsoi, 2005] Tsoi, W. C. T. (2005). The effects of occurrence frequencies on second language acquisition: A quantative comparison of cantonese, mandarin, italian, german and american english. Chinese University of Hong Kong.

[Vallabha and McClelland, 2007] Vallabha, G. K. and McClelland, J. L. (2007). Success and failure of new speech category learning in adulthood: Consequences of learned hebian attractors in topograhic maps. *Cognitive, Affective, & Behavioral Neuroscience*, 7(1):53–73.

[Vanvik, 1985] Vanvik, A. (1985). *Norsk Uttaleordbok: A Norwegian pronouncing dictionary*. Universitetet i Oslo.

[Wallace, 2006] Wallace, A. (2006). D'arcy Thompson and the theory of transformations. *Nature Review Genetics*, 7:401–406.

[Wang et al., 2012] Wang, R., Perreau-Guimaraes, Carvalhase, C., and Suppes, P. (2012). Using phase to recognize english phonemes and their distinctive features in the brain. *PNAS*, 109(50):20685–20690.

[Wells, 1982] Wells, J. C. (1982). *Accents of English 1*, chapter The reference accents. Cambridge University Press.

[Werker, 1989] Werker, J. (1989). Becoming a native listener. *American Scientist*, 77:54–59.

[Wilson and McNaughton, 1993] Wilson, M. A. and McNaughton, B. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, 261(5124):1055–1058.

[Yartsev et al., 2011] Yartsev, M. M., Witter, M. P., and Ulanovsky, N. (2011). Grid cells without theta oscillations in the entorhinal cortex of bats. *Nature*, 479(7371):103–107.

[Yildiz, 2005] Yildiz, Y. (2005). The acquisition of english interdentals by turkish learners: explaining age effects in l2 phonology. Paper presented at 10th Essex Graduate Conference in Linguistics, University of Essex.

[Zwicker, 1961] Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*, 33(2):248–248.