Perceiving whistled speech : a study of musicians' capacity for language processing

Anaïs Tran Ngoc

Perception de la parole sifflée: étude de la capacité de traitement langagier des musiciens

Anaïs TRAN NGOC

Laboratoire Bases Corpus et Langage

Présentée en vue de l'obtention du grade de docteur en Sciences du Langage **de l'**Université Côte d'Azur

Dirigée par : Fanny Meunier
Co-dirigée par : Julien Meyer
Soutenue le : 11 décembre 2023

Devant le jury, composé de :
Rachid Ridouane, Hdr, CNRS, Université Sorbonne Nouvelle Paris – 3
Daniele Schön, Hdr, CNRS, Institut de Neurosciences des Systèmes
Diana Passino, PR, Université Côte d'Azur
Léo Varnet, Chargé de Recherche, CNRS

Perceiving whistled speech: a study of musicians' capacity for language processing

Jury :
Rapporteurs
Daniele Schön, Hdr, CNRS, Institut de Neurosciences des Systèmes
Rachid Ridouane, Hdr, CNRS, Université Sorbonne Nouvelle Paris – 3
Examinateurs
Diana Passino, PR, Université Côte d'Azur - Président du jury
Léo Varnet, Chargé de Recherche, CNRS
Directeurs de Recherche
Fanny Meunier, Hdr, CNRS, Laboratoire BCL, Université Côte d'Azur
Julien Meyer, CR, CNRS, Gipsa-Lab, Université Grenoble Alpes

Perception de la parole sifflée: étude de la capacité de traitement langagier des musiciens

La perception de la parole est un processus qui doit s'adapter à un grand nombre de facteurs de variabilité. Ces variations, qui modifient le signal sonore, incluent des spécificités de production chez les locuteurs. En utilisant un signal de parole modifiée de manière expérimentale, nous pouvons cibler certains aspects du signal, pour mieux comprendre leurs rôles dans les processus perceptifs. Dans cette thèse, nous traitons une forme de parole naturellement modifiée, appelée « parole sifflée », pour explorer le rôle que jouent les indices phonologiques lors de la perception de la parole. Cependant, ces facteurs de variabilité concernent également la réception du signal, où l'écoute est influencée par l'expérience de chacun. Nous nous intéressons ici à l'effet de la pratique musicale classique sur la perception de la parole sifflée.

La parole sifflée augmente le signal de la parole modale vers le registre de fréquences le mieux perçu par l'oreille humaine. Dans notre corpus, les voyelles se réduisent à des fréquences sifflées dans un registre propre à chaque voyelle, et les consonnes modifient ces fréquences selon leur articulation. Dans un premier temps, nous avons considéré la manière dont la parole sifflée est traitée par des personnes n'ayant jamais entendu ce mode de parole auparavant (écouteurs naïfs). Nous avons considéré quatre voyelles et quatre consonnes cible : /i,e,a,o/ et /k,p,s,t/, analysées dans un contexte isolé et dans la forme VCV, ainsi que dans des mots sifflés (choisis pour intégrer ces mêmes phonèmes). Nous avons ensuite considéré l'effet de la pratique musicale sur la perception de la parole sifflée, en nous intéressant également à différentes façades de l'impact de la pratique musicale : le type de traitement, le transfert de connaissance et l'effet du niveau et de l'instrument d'apprentissage.

Les résultats montrent que tous les auditeurs catégorisent les phonèmes et les mots bien au-dessus du hasard, avec une préférence pour certaines caractéristiques acoustiques, soit des phonèmes (consonnes ou voyelles) ayant des contrastes de fréquence. Cette facilité est néanmoins affectée par le contexte du phonème (notamment dans le mot). Nous observons dans un second temps un effet de pratique musicale continue selon la quantité d'expérience, mais qui est d'autant plus marqué pour des personnes avec un haut niveau de pratique. Nous attribuons cet « avantage » musical à une meilleure exploitation d'indices acoustiques, permettant un transfert de connaissances musicales vers la parole sifflée, bien que l'effet de transfert reste inférieur à une expérience de pratique sifflée. Cette exploitation acoustique est spécifique à l'instrument pratiqué, avec un avantage marqué pour les flûtistes, surtout dans le traitement des consonnes. Ainsi, l'effet d'un entraînement, tel que la musique, améliore la performance selon la similarité du signal sonore d'un point de vue acoustique et articulatoire.

Mots clés : Perception de la parole, Parole sifflée, Psycholinguistique, Musique, Phonologie

---

Perceiving whistled speech: a study of musicians' capacity for language processing

---

Speech perception is a process that must adapt to a large amount of variability. These variations, including differences in production that depend on the speaker, modify the speech signal. By then using this modified speech signal in experimental studies, we can target certain aspects of speech and their role in the perceptive process. In this thesis, I considered a form of naturally modified speech known as "whistled speech" to further explore the role of acoustic phonological cues in the speech perception process. Variation, however, is not unique to speech production: it is also present among those perceiving speech and varies according to individual experience. Here, I analyzed the effect of classical music expertise on whistled speech perception.

Whistled speech augments the modal (spoken) speech signal into higher frequencies corresponding to the register best perceived by human hearing. In our corpus, vowels are reduced to high whistled frequencies, in a pitch range specific to each vowel, and consonants modify these frequencies according to their articulation. First, we considered how naive listeners (who have never heard whistled speech before) perceive whistled speech. We targeted four vowels and four consonants: /i,e,a,o/ and /k,p,s,t/, which we considered in isolation, in a VCV form, and in whistled words (chosen to incorporate the target phonemes). We then considered the effect of musical experience on these categorization tasks, taking an interest in the transfer of knowledge and the effect of instrument expertise.

In these studies, we observed that naive listeners categorize whistled phonemes and whistled words well over chance, with a preference for acoustic cues that characterize consonants and vowels with contrasting pitches. This preference is nonetheless affected by the context in which

the phoneme is heard (especially in the word). We also observed an effect of musical expertise on categorization, which improved with more experience and was strongest for high-level classical musicians. We attributed these differences to a better use of acoustic cues, allowing for a transfer of skills between musical knowledge and whistled speech perception, though improved performances due to musical experience are much lower than for participants with a knowledge of whistled speech. These acoustic skills were also found to be specific to the instrument played, where flute players outperformed the other instrumentalists, particularly on consonant tasks. Thus, we suggest that the effect of training, such as music, improves one's performance on whistled speech perception according to the similarities between the sound signals, both in terms of acoustics and articulation.

Keywords: Speech perception, Whistled speech, Psycholinguistics, Music, Phonology

Merci, à mes rapporteurs et mon jury pour votre patience et votre enthousiasme pour ce projet. Cette thèse m'a permis d'apprendre beaucoup plus que je ne le pensais, et j'espère, de contribuer de ma manière à la recherche.

Merci à mes deux encadrants de thèse (Fanny et Julien), qui m'ont soutenu dans ce projet de thèse malgré ces nombreux retournements de situation, des succès et des difficultés. Merci pour votre patience, vos conseils, vos relectures, et vos réflexions.
Merci de m'avoir aidé à trouver et à raconter cette histoire.

Merci à ma famille (Arden, Stéphane, Miranda), pour votre aide, vos relectures, pour avoir partagé et repartagé mes expériences, et pour avoir su quand ne pas trop parler de la thèse.

Merci à mes amis (Izia, Marie-Julie, Milena, Emilia, Marine, Adrien, Robin, Alex, Oonagh… et tous les autres) pour leur bienveillance au cours de ces dernières années, vos encouragements, vos relectures et pour ces beaux moments ensemble qui m'ont permis de respirer et de continuer jusqu'au bout.

Merci à mes amis co-doctorants (Nolwenn, Nico, Mar) pour avoir partagé cette expérience (les hauts, et surtout les bas) avec moi. Cette solidarité a été incroyable et je n'aurais jamais pu finir sans vous.

Finalement, merci aussi à toi, Paul, pour avoir insisté que je me repose, pour tes repas incroyables tout au long de la thèse, pour tes relectures et pour ton soutien infini !

Table of Contents

Foreword

Speech perception is a complex process involving the speaker and the listener. As speech includes variability, this perceptive process must adapt to the modifications in the signal. Variability can correspond to elements produced by the speaker, such as their accent, dialect, or the mode in which they are speaking (whispered, shouted). Other aspects of speech can also produce variability, such as the context or environment in which the speech is produced. These elements modify the speech signal by transforming or even removing speech cues, yet these changes also provide insight into the importance of such cues. Thus by testing the perception of speech variability or modifications, we can define the role of these cues in the speech perception process.

In this thesis, I considered a form of naturally modified speech known as “whistled speech”. This is because whistled speech has a particular transposition process, which replicates aspects of modal (spoken) speech and characterizes phonemes according to specific whistled cues. Thus, by further analyzing the whistled cues present and the way they are used in perception, we reveal the relationship between whistled speech and modal speech, allowing us to consider the role of such cues in speech perception more generally. However, the variability considered in production (through whistled speech) is also present in perception, where forms of individual experience can affect listeners. Elements which affect speech perception in the listener may include differences in spoken dialect, a knowledge of one or several foreign languages, or musical experience. Because of the similarity between whistled speech timbre and cue production (resembling a musical melody), whistled speech and music seem particularly similar. Thus, in this thesis, I take an interest in the effect of musical experience on whistled speech perception. In doing so, I seek to answer the

following question: how is speech perception affected by variability in the signal (through the form of whistled speech) and in the listener (due to musical experience)?

To address this question, I constructed an article-based thesis structured around seven articles and one supplementary study (four published, three under review and one in press) organized into five chapters. Chapter 1 provides a foundation for the experimental data and serves as a general introduction to the topics discussed. In it, we introduce the phoneme, our target unit in this thesis. We consider its role in speech perception, including how it is characterized and how it can be deconstructed by means of modified speech. We also introduce speech perception theories, by focusing on the relationship between perception and production. We then introduce whistled speech, the form of modified speech used to test speech perception in this thesis, before considering the effect of musical experience on speech perception (by exploring and defining musical skills). The experimental chapters (Chapters 2, 3, 4, and 5) then introduce several behavioral experiments presented by means of various articles. Chapters 2 and 3 focus on the perception of whistled speech by non-musician listeners, comparing target consonants and target vowels both in isolation and in the context of the word (for vowels) and providing insight into the effect of production variability. Chapters 4 and 5 consider the effect of musical expertise on whistled speech perception, reprising the previously described experiments by including a more diverse listening population, thus focusing on the effect of perceptual variability in the listener. A general discussion follows, which reviews the various themes presented and results obtained, and suggests openings for further studies.

Chapter 1

Introduction

1.1 Speech perception

Speech perception, a complex process that maps speech sounds onto linguistic representations, is based on the interaction between different-sized speech units. Upon receiving the spoken signal, the speech stream is segmented into different types of units (from sentences to words, to phonemes…) and reconstructed to produce meaning. In this thesis, I focus on the perception and categorization of some of the smallest units in the speech signal: the phoneme and the word. These units, though also dissectible, serve as an intermediary for processing speech on a higher level. In this first section, the role of the phoneme is described in the context of speech perception. We first justify using the phoneme in our speech perception studies by taking an interest in the controversy surrounding this unit, and it's demonstrated role in behavioral tests and word models. The abstract quality that creates such controversy also allows the phoneme to serve as an interface between larger and smaller perceptual units. Thus, after establishing the importance of the phoneme in speech perception, we shall propose a description of the phoneme according to acoustic and articulatory cues by also considering production. Finally, we will examine how the phoneme and its cues play a role in experimental studies (notably with modified speech) and how such approaches reflect speech perception theories more generally.

1.1.1 The Role of the phoneme

1.1.1.1 Defining the phoneme as a unit

Phonemes are known as mental categories of sound, or "units used to represent the psychological equivalent of a speech sound" (Baudouin de Courtenay, 1972, p. 152 as shown in Kazanina et al., 2018, p.560), where they are considered as smallest unit of sound. They are therefore, in their essence, abstract psychological representations regrouped into consonants and vowels, and expressed by speech sounds - the acoustic realizations or phonetic forms. These acoustic realizations vary according to the pronunciation context or dialectal differences, and the abstract phoneme represents and regroups each of these phonetic realizations. The sound of /k/, for example, is produced differently in "keep", [k^hip], and "actually", [æktʃuəli], however, it corresponds to a single phoneme (/k/), as these different pronunciations (aspirated or not aspirated) do not impact the meaning of the word. Replacing /k/ with a different phoneme, however, such as /p/, would change the meaning of the word. In our studies, we consider the phoneme in the context of speech perception/production as a prelexical unit, anchoring this choice in experiments that have demonstrated the importance of the phonological level. Some examples of such experiments include studies considering the segmentation of artificial languages, where listeners rely on consonant roots and groupings as tools for segmentation (Newport & Aslin, 2004; Bonatti et al., 2005; Toro et al., 2008), or dichotic listening experiments, where listeners also showed confusion due to consonant migration in CVCV words (Morais et al., 1987; Cutting & Day, 1975). These experiments underline how phonemes (or consonants in these examples) are used as an important perceptual unit. This small unit can then be used on higher levels of perception, as shown in various word models such as the Cohort model, Distributed Cohort model (DCM), TRACE Model,

Neighborhood Activation Model (NAM), Shortlist, and Adaptive Resonance Theory (ART). However, in these models various prelexical units are often included in addition to the phoneme (see McQueen 2005 for full review).

In the Cohort, Distributed Cohort, and TRACE models (Marslen-Wilson & Welsh, 1978; Gaskell & Marslen-Wilson, 1997; McClelland & Elman 1986) we observe the inclusion of features, with different definitions of the feature according to the model (the beginning of the signal in Cohort and DCM, or time-specific definitions). These features lead to a phonological representation, which then activates words. The units involved also feed backward at different levels of the process (semantic activation in DCM, and an interaction between word nodes and phonemes in TRACE). The ART model (Magnussen et al., 2012), like the TRACE model, creates chunks based on the features of the signal by activating phoneme groups before the word. The NAM (Luce & Pisoni, 1998) uses the phoneme as a pre-lexical unit, though it also relies on an acoustic-phonetic DAS (deletion, addition, subtraction) threshold to propose phoneme variation. Finally, the Shortlist model (Norris, 1994) activates phonemes whose construction feeds forward to the word options.

Thus, there is a clear presence of other prelexical units (both larger and smaller than the phoneme) involved in these word perception models, and we observe how the phoneme is not unanimously considered to be a tool that gives access to lexical selection. Indeed, possibly due to the size and abstract nature of phonemes, other units have been proposed, including spectra, features, gestures, allophones, syllables, demi-syllables, and tri-phones, among others (see Kazanina et al., 2018 for a more detailed review). One reason for this profusion of perceptual units is the difficulty in defining phonemic limits. Massaro (1974; 1975), for example, proposes that the phoneme cannot be treated as a single unit because certain consonants (most plosives) require

vowel coarticulation to produce an articulatory and acoustic form. The phoneme therefore relies on the syllabic unit to take shape. This supposes that speech perception would be based on the syllabic unit, or as Massaro (1972) also suggests, variations of it: for example the consonant-vowel (CV) or vowel-consonant (VC). Experimentally, the syllable has also proved to be an essential unit within speech segmentation (Mehler et al., 1981) and for perception more generally (Mattys & Melhorn, 2005). Healy & Cutting (1976) demonstrate that the role of the syllable is as important as that of the phoneme in a target-matching experiment, consequently proposing that both the phoneme and the syllable serve as basic units in speech perception. However, if we concur that the syllable is also considered as a prelexical unit, then other position-specific perceptive units such as the phone (Pierrehumbert, 2003), the allophone (Mitterer et al., 2018), or even the feature (Dahan & Mead, 2010) can also be considered as relevant units in speech perception.

In word perception models, notably in TRACE and the Distributed Cohort Model, these various units interact together (feed forward/feed backwards). Another example of such an interaction has been shown in an experimental context with allophones, proposed as an intermediary between features and phonemes in word perception (Mitterer et al., 2018). Mitterer & Müsseler (2013) and Reinisch et al. (2014) suggest that allophones may help access the lexicon, but that the lexicon would also help access the phoneme.

Thus, though the choice of prelexical units included in word perception models differs, the phoneme is invariably present, either as a prelexical unit or in interaction with other prelexical units, through bottom-up and/or top-down processing. This reflects the intimate and inseparable relationship between smaller prelexical units (such as the allophone or the feature), the lexical unit, and the phoneme.

In light of this, we suggest that studying the phoneme in the context of speech perception is justified, due to its role as an intermediary between processing levels. Kazanina et al. (2018) underline this idea by also highlighting the phoneme's important role in memory. Indeed, it is generally accepted that the phonological form is universally represented in long-term memory, making the phoneme an essential access code for speech comprehension and production. This may explain the phoneme's role as an interface, interacting with demi-syllables or allophones, even when larger or smaller units may be prioritized during the perceptive process (Bowers et al., 2016). This is notably possible because of the abstract nature of phonemes, which can be maneuvered to combine into syllabic or word forms. Kazanina et al. (2018) underline that theories demonstrating the role of other units, which are bigger or smaller than the phoneme (i.e. the syllable or the phone) do not discredit the value of the phoneme. As such, the presence of the phonemic unit in speech perception is supported not only through experimental demonstration but through its definition as an abstract unit, allowing it to serve as an interface between sounds, other prelexical units, and words. Therefore, we consider that using the phoneme to test and analyze speech perception processes will also give us an insight into the role of other prelexical units or cues, as well as the speech perception process more generally.

1.1.1.2 Characterizing the phoneme

Having established the importance of the phonemic unit in speech perception, we now seek to characterize it. In doing so, we find that the phoneme is central to both perception (as seen above) and production, as the phonemic unit also regroups phonetic variations (Hickok, 2014). Experimentally, the link between perception and production has been thoroughly established. Perkell et al. (2004), for example, show that contrasts in vowel production reflect participants'

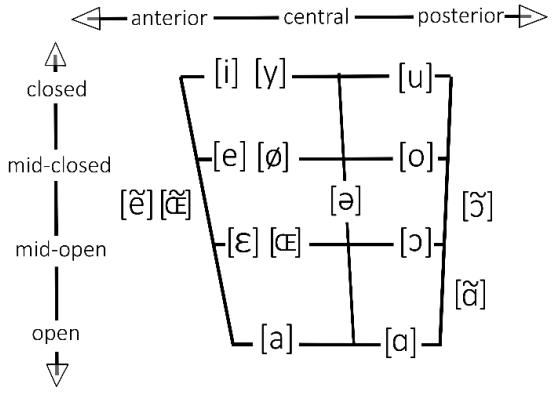
capacities for vowel perception, or, in the work of Byun & Tiede (2017) children with better perceptual acuity showed improved rhoticity in production. These results correlate more broadly with research showing a certain quantity of adaptation according to perception (Villacorta et al., 2007) which gives rise to feedback systems in production (Levelt et al., 1999; Casserly & Pisoni, 2010). As such, we can describe the phoneme in production according to both articulatory and acoustic cues, which will then enable us to further discuss speech perception theories based on these cues. We will focus on French, as it is our target language, spoken by the participants included in the experiments which we present in Chapters 2, 3, 4, and 5. However, as the whistled stimuli included are produced by Spanish speakers from the Canary Islands, we also include some phonological descriptions of the Spanish used there.

Vowels

The French language contains 12 oral vowels (including [ə]), and 4 nasal vowels. Spanish includes only 5 vowels. These two phonological systems exist within a different vowel space, though they share common vowels. In our studies, we focused on four common vowels between French and Spanish, as used in Meyer et al., 2017.

Figure 1:

French Vowel Triangle



Vowels can be characterized according to the position of the tongue in the oral cavity, which serves as a description of vowel articulation. When the tongue moves towards the front of the oral cavity, anterior vowels are formed, whereas when the tongue moves towards the back of the oral cavity, posterior vowels are formed. This first opposition places the vowels on a horizontal axis

based on the tongue's position (forward/backwards). The distance between the tongue and the palate, or the aperture, will impact how open the vowel will be. This serves as a vertical axis according to the aperture degrees: open, mid-open, mid-closed, and closed. These two axes provide a visual representation of the articulatory positions of vowels in a vocal triangle (see Figure 1).

Vowels can also be characterized according to their acoustic attributes, traditionally demonstrated using the source-filter theory (Fant, 1960; Stevens, 1998; Chiba & Kajiyama, 1958). In this theory, airflow stemming from the glottal source (which establishes the Fundamental Frequency - F0) passes through the supra-glottic filter and the vocal cavity, which changes in size to reinforce certain harmonics from the source and create formants. The articulatory elements (aperture, tongue) change the space in which the vowel is produced, thus playing a role in the formant distribution. Other aspects, such as the position of the lips and the velum can also affect the formant distribution. The acoustic characteristics (the fundamental frequency and the formants) are directly associated with articulatory elements: generally, the first formant is associated with the aperture and the second formant with the tongue position (anterior/posterior) as well as the lip shape (rounded or stretched). The third formant is also defined by the position of the lips (rounded/stretched), opposing the vowels [i] and [y] (Vaissière, 2020; Meunier, 2007). Therefore, the vowel triangle (see Figure 1) also reflects the acoustic relationship between formants 1 and 2 (F1 and F2), where plotting F1 (vertical axis) according to F2 (horizontal axis) re-creates the vocal triangle. The acoustic representation in a vocal triangle is slightly skewed however, as it does not include F3, and acoustically it may be more appropriate to consider vowels according to levels of maximal constriction (Meunier, 2007). Indeed, such constrictive points can be directly associated with certain vowels, such as [i] and [a], (Fant, 1960). Nonetheless, the vocal triangle proposes a

multifaceted characterization of each vowel, revealing a close relationship between acoustic cues and articulation in speech production.

However, both of these cues are affected by the context in which they are produced. To highlight this difference, Paillerau (2015) compares two studies on French vowel formants, one with vowels tested in isolation and the other in spoken context. Georgeton et al. (2012), tested French vowels, produced by several different speakers, in isolation. They show that F1 is generally stable for vowels, though slightly less so for [ɔ], [a], [y], [i] and [u]. In addition, the different formants vary in height according to the position of the articulatory elements, revealing defining relationships between formants, thus regrouping and chacterizing the vowels. These include the proximity between F1/F2 for [u, o, ɔ], between F2/F3 for [y], and between F3/F4 for [i]; as well as the distance between F2 and F3 for [e, ɛ]. In Gendrot & Adda-Decker (2005), vowels were extracted from word contexts found in 2 hours of radio shows. The vocal triangle constructed using the first three formants of these vowels mirrors the one produced by Georgeton et al., 2012, suggesting that both isolated and in-context vowels employ relatively stable articulatory and acoustic cues. However, the representation proposed by Gendrot & Adda-Decker (2005) highlights a more centralized vowel distribution within the space with fewer extremes than among isolated vowels, thus showing an effect of context – and coarticulation - on vowel cues. Coarticulation seeks to minimize efforts in certain registers of speech (Farnetani & Recassens, 2010) and to make speech more intelligible for listeners by maximizing differences (Scarborough & Zellou, 2013), affecting both vowels and consonants.

Acoustically, descriptions of coarticulation in vowels show that it can affect both vowel transitions (in English, Öhman, 1966) and the stable part of the vowel (in French, Durand, 1985).

Vowels, however, are known to maintain a certain acoustic "stability", as noted in the Quantal Theory (see Stevens, 1989) and observed in Gendrot & Adda-Decker (2005). This theory defines the quantal space as the zones where articulatory variation will not have a strong effect on vowels. Such zones correspond specifically to vowels with a strong constriction, notably /i/, /u/ and /a/. These "stable" vowels have been defined by the proximity between formants according to the Dispersion-Focalization theory (Shwartz et al., 1998), and are called focal vowels. Due to this proximity, these focal vowels ([i], [y], [u] and [a]) would be perceptually more stable with an increased "perceptual value".

Consonants

In French, there are 17 phonemic consonants, whereas, in Spanish, there are 19 consonants. Variations exist in consonant production according to the dialect used in both of these languages. In our studies, we focused on four consonants that are produced similarly in Spanish and French.

Consonants, like vowels, can also be characterized according to their articulatory and acoustic elements. Acoustically, several main classes of consonants can be defined according to their mode of production: plosives (or occlusives) where the vocal tract is entirely shut in production, fricatives (or constrictives) where the airflow is obstructed creating a turbulent noise, and vocalic consonants which can also contain formants (nasals can be included in this group; see Meunier, 2007). These somewhat acoustic descriptions are completed by other articulatory measures with the tongue serving as a main articulator and the inferior lip as a second source of articulation. In each of these consonant groups, the place of articulation, or where the consonant production occurs, is used to further define consonant categories: alveolar, velar, and bilabial (among others). Other acoustic cues, such as the way the airflow is obstructed (trill, fricative, and approximant) as

well as aspects such as voicing, nasalization, or aspiration can further qualify the acoustics of the consonant. For example, the bilabial articulation placement includes /p, b, m/, where /p/ is an unvoiced bilabial plosive, /b/ is a voiced bilabial plosive, and /m/ is a voiced bilabial nasal.

As voiced consonants (such as nasals and laterals) also contain formants (Fant, 1960), they can, like vowels, be characterized by specific formant frequencies (Ladefoged, 1993). For plosives, the transition between the point of articulation and the vowel creates distinguishing patterns for consonant formants. As F1 rises for all articulation points, this applies especially to F2 and F3. Velar consonants, for example, can be characterized by a "pinch" between F2 and F3. The formant transition towards the vowel, and the frequency attained at the vowel, can be called the "locus" or "locus frequency" (Hewlett & Beck, 2010). Similarly to vowels, consonants are also affected by coarticulation. These modifications are both local (directly influencing the surrounding phonemes) and far reaching. This includes the effect of a consonant on a vowel, or the reverse, either in the order heard or through anticipatory effects (Öhman, 1966; Recasens, 1987, Farnetani & Recasens, 1993). Coarticulation also extends to anticipatory vowel-to-vowel coarticulation, for example in the context of a VCV segment, or consonant-to-consonant anticipatory effects (CVC) (see Feng et al., 2011). These coarticulation effects between vowel and consonant productions can also affect word perception (see Nguyen, 2001).

As the main phonological variations of Canary Island Spanish generally affect the consonants rather than the vowels, we can also consider consonant variations in this dialect of Spanish used by the whistlers tested in this thesis. The official language of the Canary Islands is Castilian, though in reality, several different Ibero-roman dialectal variations are spoken in this archipelago. They have been grouped under several terminologies: Atlantic Spanish, Meridional Spanish, or more

commonly “Canarian” (*canario* in Spanish). Consonant variations specific to the Canary Islands include an aspirated syllable-final /s/ (non-aspirated in El Hierro Island at the end of a speech group, Díaz, 2008) and a lack of distinction between /s/ and /θ/ (also common in Andalusia). In addition, characteristics such as a velarized phrase-final consonant, a prevocalic /n/, the retention of the liquid /ɦ/, the reduction of syllable/word-final liquids, and an erased intervocalic /d/ (Lipski, n.d.; Alvar, 1955), have also been noted (see Table 1). In our studies, we focused on four common consonants between French and Spanish (/k/, /p/, /t/ and /s/). Among these consonants, only the /s/ showed dialectal variations specific to the Canary Islands, however we generally avoided ambiguities (final /s/, or /s/ and /θ/ similarities) in our choice of stimuli.

Table 1:

Consonant groups of the Canary Islands (see Brós et al., 2021)

	Bilabial		Dental		Alveolar	Palatal		Velar	
Nasal	m				n	ɲ			
Plosive	p	b/β	t	d/ ð		ʧ	j	k	g / ɣ
Fricative	f				s	ʃ		x/h	
Lateral					l	ʎ			
Flap					ɾ				
Trill					r				

1.1.2 Testing speech perception

1.1.2.1 Using synthetic speech to understand perception

Due to the complexity of the speech perception process and the integration of so many smaller units and cues within each phoneme, it is difficult to understand the role each element plays in speech perception. To overcome this obstacle, researchers have sought to deconstruct and reconstruct speech: by using the least amount of cues possible, one may establish which acoustic cues are relevant to the perception of the segment (Pisoni, 1979). The creation of synthetic speech, using minimal acoustic tools coupled with perceptual tests, provided the first insight into the acoustic makeup of phonemes. Indeed, the creation of synthetic phonemes using Pattern Playback (converting hand-drawn formants into sound) has helped define the make-up of vowel sounds and the importance of certain cues. This method relied on the proximity between certain formants to produce "simplified" vowels: reducing close formant pairs to a single formant and anterior vowels to two formants. Several perceptive tests showed that despite these reductions, vowels were well recognized (Delattre et al., 1952). The one-formant vowel reductions included [o], where F1 and F2 were assimilated to one formant close to F1, [ɔ] where F1 and F2 were assimilated to a formant in between the two, and [i], where F3 and F4 were reduced to an intermediate formant. These synthesized vowel formants are also known as "focal" vowels – or F2' (see Carlson et al., 1974). Chistovich and Lublinskaya (1979) show that this value can be calculated when the formants are close enough to each other (by 3-3.5 Bark). This formant peak, which corresponds to the tonotopic distance between close formants, is also known as the center of gravity. However, a large downfall in this reduction was the lack of representation in variability between speakers, which Strange (1989) calls the "speaker normalization" problem.

To compensate for such problems, Ladefoged & Broadbent (1957) suggest that listeners calibrate the vowel space according to the speaker. To do so, Strange (1989) suggested taking into consideration the non-linear relationship in human sound perception (known as the "Elaborated" model). Categories are therefore established between F0-F1, F1-F2, and F2-F3, and vowels are classified according to whether each dimension exceeds the 3-3.5 Bark critical difference. By using the Bark scale (Syrdal & Gopal, 1986), phonological differences due to age or gender do not affect vowel measures. Such discussions and representations, however, have focused on isolated vowels, or vowels in the CVC structure. The Dynamic Specification approach thus further investigated this question of variability in vowels, by considering the role of the vowel context (Strange, 1989).

Like the vowels, consonants have also been recreated through synthetic productions, which, by being able to control acoustic parameters, has helped develop our understanding of acoustic cues in consonant perception. The first experiments using synthetic consonants considered the relationship between noise bursts and plosive consonants (/p/, /t/, and /k/), showing a distinction between these consonants based on the frequency of the burst, as well as on the vowel with which it was paired (Liberman et al., 1952). Another study considered the role of formant transitions in the perception of stop consonants. This showed that the 1st formant provided voicing and manner cues and the 2nd formant provided articulation cues in stop consonants (Cooper et al., 1952). In Liberman et al. (1954), consonant-vowel transitions were recreated for /b/, /d/, /g/, /p/, /t/ and /k/ using painted representations of the spectrogram with a focus on the F2. This consonant categorization task opposed "plus" frequency shift modulations (rising above the F2), heard as /t/, /d/ or /g/, /k/, and "minus" transitions heard as /p/ or /b/. These categorization types also depended on the coarticulation with the following vowel: /pa/, /ta/, and /ka/, for example, could be

distinguished from one another, but /ki/ and /ti/ could not. This was also the case for fricative consonants.

However, although studies using formant-based reductions help identify certain cues necessary for speech perception, Duffy & Pisoni (1992) showed that these minimal cues can easily be missed when they are not supported by other redundant cues. Indeed, in Clark et al. (1985), participants continuously improved when identifying normal speech syllables in noise by using other cues to increase their performance. This was not the case for synthesized speech with fewer speech cues. This suggests that though synthetic speech may recreate certain speech cues, the reduction of phonemes to only one or two acoustic cues can be misleading, possibly because it is lacking the redundancy present in normal speech. In the context of synthetic words in isolation, listeners also show difficulties identifying phonemes in isolated words, reflected in the amount of time needed to perform the task as well as the accuracy ratings (Duffy & Pisoni, 1992). Results show improvements for sentences in synthetic speech, which increased by 14% compared to isolated words (Mirenda & Beukelman, 1987), suggesting the importance of additional cues (such as top-down processes). These studies suggest that synthetic speech can be a useful tool for understanding speech perception, notably by improving our understanding of formants in vowel and in consonant perception, however, this may be reductive of the perceptive process in modal speech.

1.1.2.2 Speech perception theories

Acoustic and auditory theories surrounding speech perception are a reflection of synthetic speech and acoustic-based findings. In classic acoustic theory, perception primitives are similar to the spectral components of synthetic speech, which serve as linguistic features for units such as phonemes or words. As described in the previous section, the search for invariance in the spoken

signal while facing the variability of speech (Fant, 1960), leads to the representation of acoustic features in patterns, where perception relies on certain "stable" zones, such as the center of gravity, quantal theory (Stevens, 1989), or according to the locus – the point within the frequency space which corresponds to the place of articulation. These auditory approaches towards perception focus only on the acoustic signal and its link with linguistic units of perception, without considering production: in acoustic theories, the link with production can be found within one's linguistic representation (through features for example; Stevens, 1998).

Such approaches oppose those of Liberman et al., 1967, who propose that speech perception is not based on invariance in acoustic representation, but rather invariance in articulation. The objects of speech perception are therefore the intended phonetic gestures produced through invariant motor commands. The listener perceives these phonetic gestures which neither correspond directly to observable articulatory movements nor to specific acoustic cues. (Liberman & Mattingly, 1985). Thus, in this Motor Theory, production and perception are inherently linked, and Liberman & Mattingly (1985) even suggest that this link may be biologically based, where an adaptive function specific to language processing would directly and automatically convert the acoustic signal into gestures.

The opposition between these two speech perception theories has however been put into question, as more recently, theories have been proposed that consider both articulatory and acoustic aspects in perception. One example of such a theory is the Perceptuo-Motor Theory (Schwartz et al., 2012), which proposes that speech perception relies on the interaction between the production of speech (based on articulatory gestures) and perceptual elements (acoustic, but

also visual). In light of these theories, the role of the cues used in synthetic speech are considered differently.

1.1.2.3 Modified speech as a tool for speech perception

Similarly to synthetic speech, modified speech is another way of studying phonological boundaries, of considering the acoustic and articulatory cues present within the signal, and of applying speech perception theories. Variations (or modifications) in the speech produced by the speaker (including accents, and dialectal variations), or by elements that are external to the speaker (for example noise, or an acoustic filter), also provide more ecological testing conditions of speech. As these modifications conserve only certain elements of the speech signal, therefore reducing the perceptual elements, the perceptual process can be analyzed specifically according to the cues present. Some examples of such studies include those focusing on noise or on foreign language perception.

Noise is often tested because it is omnipresent in various daily circumstances, thus constantly affecting speech perception. One example of this is Alwan et al. (2011)'s study, which measured the perceptual cues of syllable-initial plosives, fricatives, and alveolar/labial pairs in different levels of noise. This study, focusing on acoustic cues, showed the importance of formant frequencies, spectral amplitude, and burst/noise duration. Using noise as a tool can also incorporate variability in the speakers, processing differences between L1 (native speakers) and L2 speakers (second language speakers). Rammell et al. (2019), for example, showed that in noise L2 listeners use a bottom-up listening strategy, whereas L1 listeners use a top-down strategy.

Another form of natural variation used to explore speech perception processes is foreign language, where studies often focus on phonological categorization to define the boundaries of the phoneme within a language. This is because the link with one's native tongue is apparent from early childhood, where infants warp their initial perceptual space according to their L1 (Kuhl et al., 2008), thus affecting the perception and production of foreign speech. Behavioral tests based on such differences can give insight into the acoustic or articulatory cues used by the native speaker. For example, Japanese speakers have difficulty perceiving the differences between English /r/ and /l/, as the Japanese language only contains a rhotic flap /ɾ/ (Goto, 1971; Miyawaki et al., 1975). This obstacle can be attributed to a different distribution of the first (F1), second (F2), and third (F3) formants (Lotto et al., 2004) between the phonemes. More generally, the effect of an L1 filter when applied to the L2 can be interpreted through several different phonological integration models that characterize the relationship between the L1 sounds and the L2 sounds. Two models describe this filtering process: the Perceptual Assimilation Model (Best, 1995) and the Speech Learning Model (Flege et al., 1997). In the first model (Best, 1995), the degree of similarity between two languages will determine the difficulty speakers have in perceiving L2 sounds which will be categorized as either similar or different. In the second model, the L2 sounds are compared to the L1 sounds and placed in categories such as identical, different, or similar (in their sub-categories depending on the entrenchment of the L1 categories of the learner).

Native language can therefore be considered as a form of expertise that affects perception. However, the impact of expertise on speech cues also extends to other contexts: in a study that compared the perception of synthetic speech and normal speech, Simpson (1975) showed that relevant knowledge, such as being familiar with specialized vocabulary, affected performances.

Indeed, in Simpson's study, pilots and non-pilots heard aircraft messages in normal and synthesized speech. The pilots performed better than non-pilots and this advantage was even larger for synthetic speech. This suggests that condition-specific knowledge can affect speech perception of modified speech, where fewer speech cues may be present (Duffy & Pisoni; 1992). Thus, by testing both listener expertise and speaker variability, we are able to gain insight into the relevance of speech cues as well as to develop our understanding of the natural variation in the perceptive process.

As established in this section, speech perception is an intricate process within which different smaller units (phonemes, features..) interact in order to perceive larger units, such as words. We choose to focus on the phoneme, which, despite and because of its abstractedness, generally serves as an interface between perceptual levels. The phoneme can be described according to both articulatory and acoustic cues. However, to understand the characteristics of these units and the role these cues play in perception, research has relied on other forms of speech such as synthetic or modified speech. Findings from such speech forms have produced several speech perception theories, where acoustic and articulatory cues often serve contrasting roles in perception. However, using forms of modified speech to understand speech perception reprises the advantages of synthetic speech, by reducing or changing specific elements in the signal all while maintaining both the articulatory and acoustic cues as well as the natural variation. In this thesis, I therefore consider a modified speech form known as "Whistled Speech" as a tool to understand speech perception.

1.2 The Case of whistled speech

In this section, we present whistled speech, a modified speech form that we will use as a tool to test speech perception. We first present whistled speech in the world, developing upon the whistled transpositions and language types. We then give an overview of Silbo, the form of whistled Spanish that we used as stimuli, describing vowel and consonant production. Finally, we provide a description of previous perceptual studies on whistled speech.

1.2.1 Whistled speech in the world

1.2.1.1 Introducing whistled speech

Whistled speech, a form of naturally modified speech, uses a reduced amount of phonetic and phonological cues to construct whistled words and whistled sentences, all while maintaining natural speech communication. Indeed, this unique speech form preserves the lexicon and the syntax of the non-whistled language, which makes it a perfect tool for exploring the limits of speech perception as well as the role of phonological features and cues.

Developed in regions with extreme topography, for example, rugged terrain, mountains, or areas with dense vegetation, whistled speech serves as a solution for areas where communication is challenging. In such conditions, speaking or shouting is extremely difficult. For example, when shouting in a dense forest at a distance of 90 meters, the word recognition rate is 75%. This quickly diminishes once the distance is increased to 150 meters, requiring speakers to lengthen their sentences. At a distance greater than 200 meters, communication becomes almost impossible, as speakers start suffering from voice damage (Meyer et al., 2018). Generally, shouting in these

conditions causes an increase in fundamental frequency and amplitude. In whistled speech, by contrast, the modal acoustic characteristics are augmented to higher frequencies according to the whistler's voice and skill. This allows for the signal to be carried further, easily surpassing speaking or shouting: reaching 100m in cloudy mountains, 1 km in open mountainous valleys, and 8 km in exceptional sound-propagation conditions, depending on the whistling technique, (Busnel & Classe, 1976). As the geographical conditions described are the only criteria for the development of whistled speech, its use is not restricted to a single language or location. Different whistled languages have thus been found on all continents and in a great diversity of languages.

1.2.1.2 Historical representation

Historically, forms of human acoustic communications akin to whistled speech are mentioned as early as the 2nd century, notably in Greek texts describing whistles used in North Africa. In the Canary Islands, the first mention of whistled speech was possibly in 1402 when two Franciscan monks recounted their expedition with Jean de Béthencourt to the Canary Islands. They describe a "strange and particular language" as practiced by the indigenous Berber populations generally called the Guanches (Busnel & Classe, 1976). However, it is only in the 19th century that whistled speech is explicitly mentioned, after the indigenous Berber languages of the Canary Islands were no longer in use (due to the imposed use of Spanish through colonization). At the time, whistled speech was considered "an entertaining phenomenon" resembling both music and prosody. Quedenfeldt (1887), who described whistled languages of La Gomera Island in the Canaries, even proposes that musicians would be better suited to understand whistled speech, as he notes that whistled speech contains pitch variations. This idea was pushed further when Quedenfeldt asked musicians to transcribe whistled speech into notes. These were then whistled

back to Gomero whistlers, however, it was impossible for them to understand the original message. According to Busnel and Classe (1976), the ultimate failure of this small experiment was due to the extremely strong influence of the Western musical system, which forced the musicians to reproduce a "classical" transcription. Yet, when considering these transcriptions today, we note that the phrase was transcribed with notes corresponding to both specific vowel pitches (the /a/ as a G4) and certain consonant movements (the /s/) which are coherent with the whistled language. However, the alignment of the phrase and the notes is not coherent, and transcription lacks essential articulatory cues, which may explain why understanding the "whistled" version was impossible (see Figure 2.6, p.19 in Meyer, 2015). This early suggestion nonetheless shows an intuitive link between whistled speech and music, as well as the difficulties a listener would have in understanding whistled speech, especially if they had never heard it before. Indeed, we note that the lengthening of certain speech sounds over several "notes" is particularly complex.

The first linguistic analyses of whistled speech can be attributed to Cowan (1948), whose description of a whistled language used by the Mazatec Indians in Mexico changed the probationary analytical approach of 19th century descriptions. Contrary to the first articles on the whistled speech of the Canary Islands, he links the Mazatec whistled speech to the linguistic characteristics of the spoken language, calling it "parallel to spoken conversation as a means of communication", and employs detailed ethnographic descriptions of its use in society to decode its features. This description led to several other linguistic analyses of whistled speech, refining descriptions of whistled speech on the Canary Islands (Classe, 1956; 1957), as well as in other regions around the world. Cowan's findings (1948) also suggest that whistled speech is a natural extension of the spoken modality, leading to a more detailed list of whistled languages around the world.

1.2.1.3 Whistled language typology

Due to the diversity of whistled languages, typological categorizations of these languages reprise the tonal and non-tonal language groupings. Indeed, Busnel & Classe (1976), contrasted tonal languages with another family of non-tonal or articulated languages, and Rialland (2005), proposed a classification grouping languages based on their transposition strategy ('formant-based whistling' for non-tonal languages and 'pitch-based whistling' for tonal languages). Meyer (2005; 2015) further developed these whistled transposition-based groups by including more languages and by describing the diversity of whistled speech in more detail. These whistled typologies thus maintain the same groupings as those of tonal and non-tonal languages in modal speech, illustrating a way in which whistling strategies reflect language characteristics.

There are three main language groups: the non-tonal whistled languages, the tonal whistled languages, and an intermediate group. In non-tonal languages (such as Spanish or Greek), the vowel qualities are transposed into whistled pitches, modified by consonant articulations. Tonal languages (such as Hmong and Mazatec) base the whistled transposition on the fundamental frequency of the voice, producing a whistled form through direct association. Finally, the intermediate group is timbre-based, with a shared influence of both the fundamental frequency and formants (Meyer, 2015).

Our main focus will be on the non-tonal whistled language group, established by Busnel & Classe (1976), Rialland (2005), and Meyer (2005; 2015), as we further detail the whistled transposition of phonemes in our studies. In the modal form of non-tonal languages, the spoken pitch and pitch variation carry very little semantic information, as they produce the vowel pitches according to sentence intonation. In their whistled form, the modal vowel qualities (most notably

the formants) establish different pitches or frequency ranges within which the whistled vowels are produced. These specific whistled frequencies, which correspond to each vowel, are therefore the result of an approximation of the natural vocal tract articulation in the spoken version of the vowel. Thus, with the added constraint of an almost closed mouth needed to produce the whistle, there is an emphasis on the upper resonant cavities, producing a whistled pitch. This can cause the whistled frequencies to reflect frequency shapes of modal speech, often corresponding the F2, though sometimes the F3 for front vowels (Shadle, 1983; Meyer, 2015). Different vowel types are whistled at different frequency levels and within a certain degree of variation (mainly due to coarticulation). In the case of coarticulation between back vowels and velar/uvular consonants, whistled F0 also often resembles the F1 of modal speech.

Consonants are formed using the articulation present in spoken words, and among non-tonal languages they primarily reflect the shape of the formants which modulate the vowel frequencies. These formant shapes resemble those observed in some spoken formant transitions, notably towards acoustic consonant loci between vowel pitches. Thus, the transposition into whistled speech changes the role of pitch from supra-segmental to phonological, modifying the significance of certain cues, as well as the general perceptual process. Hyper-articulation is often necessary to compensate for the difficulty in sound production with whistled speech, and therefore, the speed of whistled sentences is slightly slower than spoken speech. Despite language-specific differences in vowel and consonant productions, non-tonal whistled languages in this group (including Spanish, Greek, Turkish, Béarnese, and Tamazight), share such transposition strategies (Meyer, 2015). Yet, these speech forms also maintain some of the same variability present in modal speech, such as inter-speaker differences and dialectal variation, principally when these differences affect

consonant articulation (which is the case between the whistled Spanish of the island of El Hierro and of La Gomera; Díaz, 2008).

1.2.1.4 Whistled speech production

To produce the sound, whistled production strategies are chosen according to the distance between the whistlers, the type of whistling activity, and the topography. For short, flat distances (up to 50 meters in a quiet environment), bilabial whistling is most common. For medium or long distances, a linguo-dental form is used where the tongue is retroflexed against the lower incisors, using either one or two fingers. The "one-finger" method uses a curved finger that presses either on the whistler's tongue (inside the mouth) or on a blade of grass, forcing the air over the middle knuckle (in a "V" shape) and out the small opening between the lips, finger and front teeth. When two fingers are used, they also form a "V" on the tongue. Other hand techniques used in dense vegetation include creating a cavity with one's hand, which can produce a larger frequency bandwidth (Meyer, 2015). The type of technique used can also influence the whistled pronunciation. Indeed, in a formant-based whistled language such as Spanish, certain techniques can make consonants harder to pronounce, even though all whistlers try to articulate whistled speech as closely to modal speech as possible. For example, when using the one-finger technique in the La Gomera dialect, labials can be harder to produce, as the finger rests upon the tongue. In our studies, we included productions from two whistlers who both used the same one-finger technique.

1.2.2 Silbo and whistled speech perception

1.2.2.1 Silbo whistled language and phonological system

In the experiments proposed here, all of the whistled stimuli used were produced by Silbo whistlers. Silbo or the whistled language used in the Canary Islands (an archipelago made up of seven main islands - Tenerife, Fuerteventura, Gran Canaria, Lanzarote, La Palma, La Gomera, and El Hierro), was, in the past, extensively practiced. Today, only a few traditional whistlers remain, notably in La Gomera, El Hierro, Gran Canaria, and Tenerife (Díaz, 2008, Meyer & Díaz, 2017). Whistled speech was preserved best in the two smallest and most western islands, La Gomera and El Hierro.

Vowel transposition

As Silbo is a non-tonal language with a formant-based transposition, vowels embody certain formants in modal speech. However, researchers have observed that, due to the constraint of whistling while articulating the vowels, several vocalic reductions exist in the whistled modality. Indeed, among the five Spanish vowels (/i/, /e/, /a/, /o/, /u/), several different vowel groupings have been proposed. Trujillo (1978) went as far as proposing his own whistled phonological system for Silbo, with only two 'whistled vowels': high (grouping /i/ and /e/) and low (grouping /a, o, u/). Though this may have been a gross reduction, Classe (1957) and Busnel & Classe (1976) also suggested that the whistled /o/ and /u/ were produced at similar frequencies, and could be grouped together, thus reducing the number of vowel sounds. Studies by Rialland (2005), Meyer (2008), and Díaz (2008) therefore reprise this idea, suggesting that there are four vowel groups in production, ordered from highest to lowest according to their characteristic frequency distribution: /i/, /e/, /a/

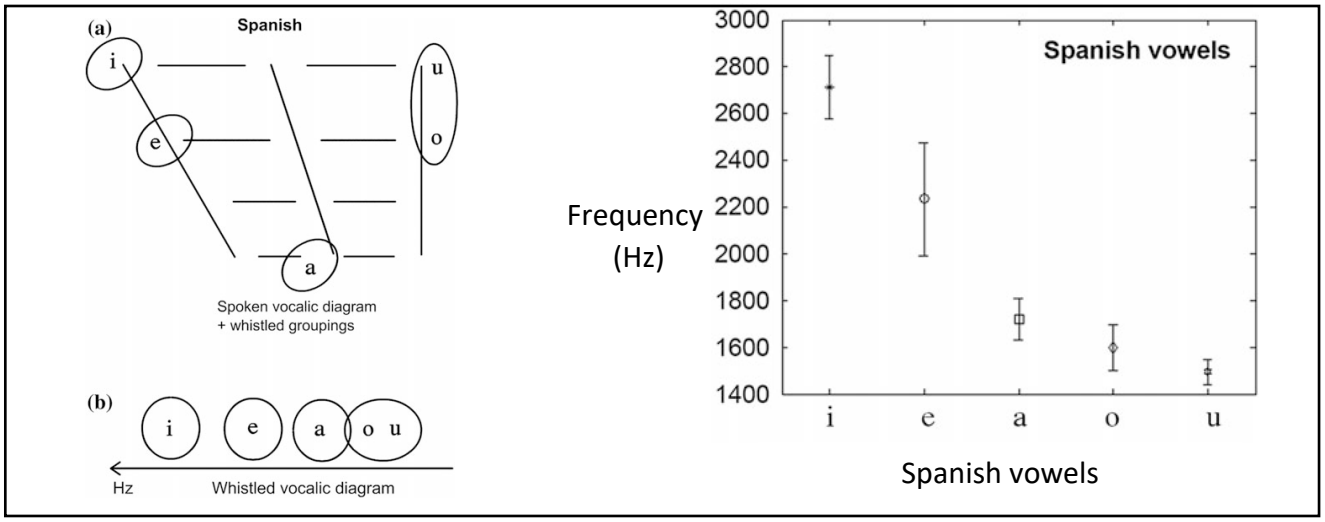
and /o, u/. Rialland (2005) gives an example of this distribution, citing the productions of a single whistler, where /i/ had an average frequency of 2620 Hz, /e/ of 1930 Hz, /a/ of 1480 Hz, and /o/ of 1380 Hz and /u/ of 1270 Hz. Indeed, the proximity between /o/ and /u/ is apparent, as they are produced at different frequencies but with a large overlap between ranges, and do not necessarily reflect a single phonological category. This proximity in frequency has since been shown to be a whistled exaggeration of pronunciation similarities present in the dialect of Spanish studied (notably the La Gomera dialect), as [o] and [u] are pronounced similarly. Meyer (2015) therefore suggests that the similarity in frequencies is a byproduct of the articulatory position, and that whistlers make use of all of the vowel phonemes.

The decrease in approximate frequency for each of these vowels (starting with /i/) is similar to the distribution of F2 in modal vowels and reflects the anterior/posterior axis of the common vowel diagram, i.e., the more the tongue moves towards the front of the oral cavity, the higher the whistled vowel is (see Figure 2). Like the F2, the frequency of the vowels /i/, /e/, /a/ and /o, u/ gradually lowers, though the pitches do not reflect the value of the F2 (see Chpt.1.2.1.3). These pitches nonetheless create relative intervals which in-turn may help with perception, by allowing the listener to place the pitches within a whistled register. Meyer (2015) and Rialland (2005) describe how whistlers anchor their productions within a whistled range, by using a ”musical” tuning note (the /a/ according to Rialland, 2005). This underlines how approximate the representation is of both the frequencies and the relative intervals, as they each depend on individual whistler variation. Such variation is due to the whistler’s vocal tract, their skills and technique, as well as communication distance (Meyer, 2015). The frequencies of these vowels can also vary slightly according to coarticulation with different consonants. This is consistent with the

effects of coarticulation in modal speech and coherent with whistlers' descriptions of whistled production, where they describe simply pronouncing the spoken word through a whistled medium.

Figure 2

Vowel diagram and corresponding frequency height in one whistler (borrowed from Meyer, 2015, p.111 and p. 107)



Interestingly, and despite this variability, the whistled range is generally consistent in all non-tonal whistled languages (lying between 4 kHz and 0.8 kHz), and the relative distribution of vowel types is similar across languages regardless of frequency changes due to language-specific pronunciation differences. Such pronunciation differences can cause other vowel frequencies to overlap, creating different vowel groupings. In Greek for example, the whistled frequencies used for [u] overlap with [ɛ], while the whistled transposition of [ɔ] overlaps with [a], creating 3 vowel groups: /i/, /u, ɛ/, and /a,ɔ/. In Turkish, there are 4 vowel groups: /i/, /Y, ɨ/,/ɛ,œ,ʊ/, /a, o/, however, the general frequency distribution with [i] highest and [o] lowest is maintained (Meyer, 2015). In Greek or Turkish, the [u] is regrouped according to pronunciation, or more specifically the place of the 2nd

or 3rd formant (Meyer, 2015). According to Meyer, this reflects elements already exploited in vowel perception such as the center of gravity effect (according to the 3-3.5 bark limit).

Consonant Transposition in Silbo

In whistled speech, consonant productions modify the whistled vowel pitches according to the articulatory elements in modal speech, creating stops and pitch changes. However, due to the production methods of whistled speech (notably fingers placed within the mouth), it can be difficult to understand and properly characterize the articulatory cues present. This is despite Busnel's (1968) attempt to take X-rays of whistled productions, which showed tongue displacement during productions (even with the two-finger method, see Busnel & Classe; 1976). Thus, we generally rely on acoustic parameters in whistled speech to characterize consonants. Rialland (2005) proposes using the signal envelope, as the envelope shape has previously been used to preserve contrasts between major classes in English, thus playing a central role in speech processing (van Tassell et al., 1987; Shannon et al., 1995). The characterization of these envelope modulations in whistled speech (continuous and interrupted), provides the frame for using pitch changes to describe consonants. These are defined by consonant loci: "acute" (towards a high locus, well above the frequency of /i/ - see Rialland, 2005 on dentals) and "grave" (no change in loci).

These acoustic cues have served as a basis for establishing consonant groups, or categories. Trujillo (1978) contrasted four characteristics: "acute", "grave", "continuous" and "interrupted" and proposed four consonant groups each defined by two characteristics. Rialland (2005) proposed eight groups of Spanish whistled consonants, based mostly on production and perception tests, adding "sharp" and "gradual decay" to Trujillo's terms. Rialland further nuances the acute consonants by describing loci in /s/ as higher than those in than /t/ (which also has sharper attacks). Díaz (2008)

also proposed groups of consonants corresponding to the practice observed among long-time and newer whistlers learning with the Yo Silbo association. Díaz suggests that additional characteristics could also be taken into account, though these cues may only be pronounced by certain expert whistlers, and may not be generalizable. Thus, though these groupings vary, researchers generally agree with the distinctions among whistled consonants in Spanish proposed by Trujillo: "acute", "grave", "continuous" (or semi-continuous), and "interrupted" (see Table 2). Contrary to whistled vowels, these consonant groups are specific to whistled Canary-Island Spanish, differing from the seven groups in Greek (Meyer, 2015) and ten in Turkish; based on the contrasts mentioned above, the whistling technique, and a velar/labial distinction (Rialland, 2005).

Table 2: [1]

Comparison of Consonant Groups proposed

		Interrupted		*Continuous*		
Trujillo (1978)		*Acute*	*Grave*	*Acute*	*Grave*	
		ʃ, t, s	k, p	r, ɾ, l, j, ʎ, d, n, ɲ	b, m, f, g, x	
Díaz (2008)		*Acute*	*Grave*	*Acute*	*Grave & Articulated*	*Grave*
		ʃ, t, s	k	r, ɾ, l, j, ʎ, d, n, ɲ	p, b, m, f	g, x
Rialland (2005)		*Acute*	*Grave*	*Acute*	*Grave*	
		t, s	p, k, f, x	d, z, l	b, g	
	Sharp	ʃ		j, ʎ,	m	
	Gradual Decay			ɲ		

[1] In this table only certain contrasting phonemes are retained without taking into consideration the allophonic variations present in Table 1. As such we included only /t, d, r, ɾ, l, n, s, j, ɲ, k, p, m, f, x, g, ʃ, ʎ/.

1.2.2.2 Analyzing whistled speech – behavioral experiments

Perceptual experiments based on whistled Spanish (Silbo) have sought to define perception for both accomplished whistlers and for listeners who have never heard Silbo. As these experiments are presented in more detail in the introduction of each article in the following chapters, we will simply provide a brief overview of these studies.

Native whistlers – nonsense syllables

The first behavioral experiments considering whistled speech focused on native/expert whistlers. Busnel (1962) used Bearnese whistler productions in the form of a "logatome" unit, to test Béarnese and Turkish whistlers' perception. This grouping of one or several syllables respects the phonotactics of the spoken language, thus targeting certain elements of the speech signal, such as consonants and vowels. The term "logatome" used mainly in the 1960s/1970s, is now outdated and replaced by "nonsense-syllables". Busnel's experiment (1962), which asked whistlers to identify consonant-vowel (CV) nonsense syllables, showed a very low recognition rate, with only 37-50% of syllables identified correctly. Moles (1970) also took an interest in CV perception in whistled Turkish (of Kusköy), as identified by 5 Turkish whistlers. However, the results were even lower than those obtained by Busnel (between 12.5 and 33%). Rialland (2005) conducted experiments on VCV nonsense syllables, using each of the vowels in Silbo, and tested with two Silbo whistlers standing 15m apart on a hillside. Both the correct answers and the confusions were analyzed, showing that whistlers' answers were 57% correct, highlighting better responses for certain consonants and vowels. Indeed, /p/ and /t/ were recognized at 77.7% and 68.75% respectively, and /a/ was recognized 94.4% of the time.

These relatively low categorization rates can be attributed to multiple factors, some of which are underlined in Meyer (2015). Firstly, though there is still a certain degree of recognition, whistled speech is not mutually understood from one language to the next due to different phonetic systems (for example from Bearnese to Turkish), and such results would not be equivalent to those of a whistler using the same language. In addition, the targeted nonsense syllables that were tested varied between tasks (VCV for Rialland, 2005; and CV for Busnel et al, 1962, and Moles, 1970) and included very few details concerning the distance at which these whistlers exchanged messages. Finally, only a small number of whistlers participated in these tasks: Rialland's test (2005), for example, included only one pair of whistlers.

More recently, Meyer et al. (2019) conducted an experiment on syllable recognition (/ta/, /da/, /ka/, /ga/) with Tashlhiyt Berber by whistlers. This experiment, based on a previous study conducted by Meyer on Turkish whistled speech (2007), took a particular interest in the confusions between consonants, both in whistled and spoken Tashlhiyt. The results obtained showed that participating whistlers obtained between 31% and 50% of correct answers, confirming proficiency. The mean level of success was 41.7%, where /ta/ and /ka/ were recognized best and /da/ was recognized the worst.

Native whistlers – Words and Sentences

Several perceptual tests also considered whistled word perception and whistled sentence perception. Busnel (1970) tested Turkish whistlers standing 10m away on 40-50 whistled Turkish words, obtaining between 60-75% of correct responses. These answer rates are higher than those of nonsense syllables, reflecting the fact that whistled speech is usually used to express full sentences, and more specifically sentences regarding familiar topics (farming, agriculture, or family).

Interestingly this is similar to the increase between perceptual levels observed in synthetic speech (Duffy & Pisoni, 1992). Full-sentence comprehension in whistled speech has also been tested. Moles (1970) tested sentence perception with sentences corresponding to three different difficulty levels, obtaining a percentage of correct responses proportionate to the difficulty of the sentence (70% for the first level, 55% for the second level, and 22% for the third and hardest level). This suggested the importance of set conversation topics or themes. In another experiment on sentence intelligibility (with set conversation topics) which tested Greek whistlers on relatively short sentences (up to 9 words). Comprehensibility was found to be between 95 and 100 percent, with no repetition necessary (Meyer, 2005).

Naive listeners

Finally, and more recently, behavioral experiments have been conducted with listeners who have never heard whistled speech before, known as naive listeners. The majority of these experiments have focused on vowel recognition for 4 whistled vowels in Silbo: /i, e, a, o/. In one experiment, naive Spanish listeners were compared to a native Silbo whistler who obtained 87.5% of correct responses, where /i/ was categorized best at 100%, /o/ and /e/ at 87.5%, and /a/ at 75% (Meyer, 2008).

Whistled vowel perception was also tested on naive French-speaking listeners, as these vowels are common to both Spanish and French. This task contained two versions: one where vowels were proposed in an isolated context, and another where vowels were presented for categorization at the end of a whistled sentence. Results showed that French speakers categorized these whistled vowels well over chance. In this experiment, 6 participants also indicated having musical experience, and performed better on the isolated vowel task only, obtaining 64% correct

answers compared to 54% for participants with no musical experience (Meyer, 2005; 2008). In addition, a vowel hierarchy was observed in these results: /i/ was categorized better than /o/, which was categorized better than /a/ and /e/. Meyer (2008) also observed an impact of vowel interval[1], which sometimes caused a vowel categorization shift, where a vowel was mistaken for a higher vowel, and the relative interval would be maintained for the following vowel sound according to the mistake made in the preceding one. The error was propagated until a relative frequency difference between successive vowels would allow for a re-alignment of the intervals and categories.

A second experiment on whistled vowels (Meyer et al., 2017) reprised the isolated vowel task and included naive Spanish and naive Mandarin Chinese speakers. These participants were compared to the French speakers tested in 2008 (Meyer et al., 2017). All participants - including the native Chinese speakers, who had only been speaking French for a couple - obtained scores higher than chance (the Chinese speakers obtained 43.5% of correct responses). In addition, Chinese speakers obtained the same hierarchy of individual vowel performance as naive Spanish speakers, though with a lower percentage of correct responses. The naive French speakers also obtained similar results to the naive Spanish speakers, obtaining 55% of correct answers, where /i/ was recognized best (at 78.4% of correct answers). By using common phonemes between the two languages, this suggests that the parallel between whistled speech and modal speech also reflects similar speech perception mechanisms in both languages. Listener profiles were further differentiated when considering the confusions between whistled vowels. These results showed that Spanish and French listeners were able to assimilate fixed tones more easily than Chinese

[1] The difference between vowels according to the range of frequencies into which they are transposed. The vowel interval between /o/ and /a/ is 1, between /o/ and /e/ is 2, and between /o/ and /i/ is 3.

speakers. However, the naive Spanish whistlers resemble the expert Spanish whistlers in confusion profiles.

In this section, we explored a form of modified speech known as whistled speech, first underlining the global presence of this speech form, before describing whistled Spanish (Silbo) in more detail according to the transposition from modal speech. We also briefly reviewed initial behavioral experiments on whistled speech, which will serve as a basis for the following articles. We now concentrate on the role of the listener during the perceptive process, by focusing on a specific type of listener experience: music.

1.3 Musical experience, a form of listener variability

In this final section, we consider musical experience as a form of listener variability in speech. We approach this topic by reviewing the conditions necessary for skill transfer (more specifically from music to another field). We highlight the importance of similarities between music and the task in question, and, seeking to establish the possibility of knowledge transfer between speech and music, we establish some commonalities between the two fields. We then propose a review of musical transfers towards speech, starting with neurological modifications and brain plasticity, before considering the effect of musical experience on phonemes, tone languages, and modified speech perception. Finally, we take an interest in the definition of the musician and the skills acquired through musical training. This provides insight into which musical skills may be transferred during speech perception and which speech cues may be affected.

1.3.1 Transferring skills from music to speech

1.3.1.1 Musical skill transfer

To consider musical experience as a form of listener variability that will produce an effect on speech perception, we first question whether it is possible to transfer knowledge from music to other tasks.

To transfer knowledge stemming from musical skill to another field, the task at hand must correspond to capacities closely related to musical training or music-related tasks. Miendlarzewska & Trost (2014) suggest that these can include listening/auditory skills, fine motor skills, temporal processing, and attention orientation. Indeed, studies have shown that even a short amount of musical training will improve auditory skills (see Carey et al., 2015 for a more detailed review). These skills include pitch perception of spectrally rich tones (likened to natural instruments), where

musicians are faster and more accurate when discriminating pitch changes (Tervaniemi et al., 2005), differences in chords (Koelsch et al., 1999), and differences in intervals (Zarate et al., 2012). Liang et al. (2016) show that these auditory skills also extend to frequency discrimination both in silence and in noise, where the discrimination thresholds were much lower for musicians than non-musicians.

Studies have also shown that there is a clear improvement in motor skills through musical training (Schlaug et al., 2005), however, these may be specific to the instrument. Logan (2014) highlights this instrument bias in an experiment which asks trombone players to respond to high-low movements on a joystick. To play the trombone, one moves a slide “up” (towards the face) and “down” (at an angle towards the ground). Interestingly, Logan (2014) finds that movements on the joystick contrary to those that would be played on the trombone caused delayed response times for trombonists compared to non-trombonists and non-musicians. We can therefore consider that musical skills transfer to auditory or motor skills through similar experience (which we can call a near-transfer), though when tasks are too similar this can cause interference (as seen in Logan, 2014). However, as highlighted by Carey et al. (2015), when tasks are not similar enough, for example when musicians are asked to analyze an auditory scene, musical skills produce no effect on performance. This underlines the specificity of musical knowledge.

The presence of these near-transfer advantages lead us to consider broader skills used in music. Indeed, musicians rely not only on trained auditory skills and motor skills but also on memory (of the music, of technical components, of physical movements), attention (reading music and listening to others) and executive functions (allowing for selective listening and switching between several tasks in a complex environment). Therefore, in similar memory or attention tasks, we could expect musical expertise to impact performance. Indeed, in Tierney et al. (2008), musician

participants show a difference in short-term memory spans compared to other groups of musically inexperienced subjects. Nie et al. (2022) also show how musical experience affects the executive functions of working memory (see Rodriguez-Gomez & Talero-Gutiérr, 2022 for a review of the transfer of executive functions).

Yet, executive functions and memory are not as closely related to musical training, suggesting that musical knowledge could transfer to tasks that may not be so musical (far-transfer). This, however, is a controversial topic, as findings show both support and opposition concerning far-transfers towards higher cognitive functions and intelligence. Indeed, in a meta-analysis, Sala & Gobet (2017) argue that though domain-specific skills are enhanced between music and cognitive functions, there is no far transfer (towards academic skills for example). These findings are reprised by Sala & Gobet (2020) who conducted a meta-analysis of 54 different studies where children participating in musical training are tested for cognitive or academic skills. They find no effect of musical training on these skills. However, when Bigand & Tillmann (2022) re-applied the methods to the same meta-analysis, they showed an effect of musical training on these far transfer skills, especially when reconsidering elements such as randomization and the type of skills taught to the control group. These positive results are echoed in a meta-analysis by Cooper (2020), as well as in Román-Caballero et al. (2022), who focused on the benefits of learning an instrument. We also note that through this meta-analysis, Román-Caballero et al. (2022) show only slight benefits for short-term training programs (thus nuancing the benefits suggested in other studies).

Thus, though musical skills can transfer to other non-musical activities, the proximity of the activity is particularly relevant. As speech is a complex process that relies on acoustic and articulatory cues as well as higher-level processing (memory, executive functions), we suggest first

comparing the parameters involved in both speech and music to better understand the potential for transfer between music and speech.

1.3.1.2 Similarities between music and language

In music, pitch and rhythm combine to produce melody, which will be played by different instruments creating the timbre. As both pitch and timbre are essential for the transposition towards whistled speech (and more specifically in whistled phonemes), we can compare these components in music to modal speech, as each domain uses these tools differently. This comparison then allows us to construct a relationship between speech processing and musical processing.

Pitch

Pitch is essential to music, as music (more specifically western music) is based on fixed pitches (notes) which construct scales and keys, serving as the building blocks of Western music (see Lerdahl & Jackendorf, 1983; Bernstein, 1973; and Meyer, 1973). Today, most musicians rely on Hertz standards to construct the pitch scale, where the note "A" (A4) is fixed at 440 Hz. However, instruments are not perfectly tuned to the Hertz system, thus, there is a tolerance of approximately 50 cents between pitches (Haynes & Cooke, 2001). Such pitch standards have emphasized the association between note names and fixed pitches in Western classical music, promoting the development of perfect or "absolute pitch" (the ability to give the name of a note according to the pitch heard, Ross et al., 2005). In speech, the perceived fundamental frequency is associated with the pitch. For non-tonal languages (such as French, English, or Spanish) pitch acts on a paralinguistic and prosodic level. Intonation may also play a role at a supra-syllabic level and can be influenced by the speaker's voice range, which means that there is no fixed pitch in speech. Pitch is also used in

speech perception to integrate additional paralinguistic information into the signal (for example gender), Lee & Lee (2010).

Timbre

Timbre in music, first defined by Helmholtz in 1877, is used to distinguish the changes that occur when the same note is played by different instruments. However, in music, there are several diverging definitions of timbre. The first definition, as described by Siedenberg & McAdams (2017), suggests that timbre is a contributor to source identity. The second, described by Siedenburg et al. (2019), states that timbre is a perceptual attribute (in the mind of the listener) perceived simultaneously to the F0 and corresponding to the spectrum (through the overlay of multiple harmonics, characterizing the signal according to the relative amplitude of partial tones). Traube (2015) argues that this second type of timbre also describes the aesthetic and emotional quality of sound. To define the characteristics of timbre, it is essential to compare traits in relative scales of qualification as the actualization of many acoustic differences is minimal. Timbre can be both a single and a fused auditory event: individual instruments have their specific timbre, but they also combine to create a group timbre (Siedenberg et al., 2019). Thus, using qualities such as "brightness, darkness, fullness, roughness" in addition to discrete or categorical attributes allow a better definition of these differences. The specificity of timbre can therefore encompass a large group of sounds or a family of instruments, as well as individual musician-based differences (see for example Fritz et al., 2014). Recent approaches towards representing timbre include those of Reymore (2021), whose 20-dimensional representation of different timbre *qualia* was based on responses from 243 participants with a musical background, to allow for a more objective characterization of the sounds heard.

In speech, timbre is acoustically perceived at the same time as pitch and is used on a phonological level to characterize phonemes. These qualifications can be known as “color” or “quality”, and use the physical correlates of timbre- such as the spectral envelope and the amplitude envelope/contour- to create a distinction between /ʃ/ vs. /s/ and /b/ vs. /w/ and /ʃ / vs /tʃ/, respectively. These distinctions also apply to vowels. One example of this is /å/ and /i/, two vowels that differ in timbre if spoken with the same loudness and pitch (Koelsch, 2011). This also applies to vowels with the same F0, which will be perceived as higher or lower according to vowel quality (i.e. timbre) (Stoll, 1984). As the F0 is also included in spoken timbre, modifications in pitch will generally modify the timbre, affecting speech perception and pitch processing in the brainstem (Krishnan et al., 2012). Individual differences also affect the timbre of speech through the individual distribution of acoustic energy across frequencies or phonation noise (Latinus & Belin, 2011), thus contributing to voice recognition. Such changes in the spectrum in turn affect one’s pitch perception (Kuang & Liberman, 2018).

Music and Speech Processing

In our review of some of the components that make up prosody and melody, we underline how, even though music and speech share many similarities, each of these common elements serves different purposes. This has led to research looking at the similarities between music and speech processing, notably because both speech and music require cognitive capacities such as memory, attention, and executive functions to function. Atherton et al. (2018) suggest that the two domains share common resources, notably when it comes to working memory. In Atherton et al.’s study (2018), an ABX task was conducted with different types of interference, comparing linguistic stimuli (3-letter words) and musical stimuli (3-note chords). This ABX task consisted of 3 settings: matching

(music/music or linguistic/linguistic), non-matching, (music/linguistic or linguistic/music) and no interference (linguistic/silence or music/silence). Though interference occurred the most in the matching task, some interference was also found in the non-matching setting, thus suggesting some common processing (also suggested in Williamson et al., 2010; and Semal et al., 1996). Music and speech therefore share similar tools (pitch, timbre) that construct phrases and sentences, as well as common processing. Peretz & Coltheart (2003) also propose a sound processing model with an initially common phase between music and speech. These processing models generally suggests that musical knowledge should transfer towards speech.

1.3.2 Musical experience and speech perception

In this section, we review forms of transfer observed, where musical experience can cause speech-related modifications.

1.3.2.1 Neurological modifications

The first evidence of a transfer between musical training and speech is modification in the brain, known as brain plasticity. Brain plasticity, or the ability "(for) the nervous system to change its activity in response to intrinsic or extrinsic stimuli" (Mateos-Aparicio & Rodriguez-Moreno, 2019, para 1), implies changes or modifications in the brain structure due to certain activities, often those requiring repetition or training. As musical training includes not only auditory and sensorimotor processes but also high-order cognitive functions, brain plasticity resulting from musical training

corresponds to changes in several different areas. Herholz & Zatorre (2012) even suggest that the multimodal nature of musical training may enhance plasticity.

Cross-sectional studies comparing “musicians” (with a pre-established musical experience) and non-musicians are often used to study brain plasticity, and in a recent review, Olszewska et al. (2021) highlight several anatomical changes present in the brain observed through cross-sectional studies. These include differences in the development of the temporal and frontal areas (Gaser & Schlaug, 2003) due to an increase in grey matter (in frontal areas, the hippocampus, and the lingual gyrus) or in cortical thickness (specifically in the somatosensory cortex, due to physical contact with the instrument, Bermudez et al., 2009). These increases in grey matter also correlate with changes in white matter architecture in regions linked to fine motor control and sensory processing. In addition, other data such as fMRI studies show extended activation in temporal areas, parietal areas, the frontal lobe, and primary/supplementary motor areas during passive listening (Bangert et al., 2006). Finally, in EEG/MEG studies, musician participants show a higher amplitude of brain-generated electro-physiologic or magnetic potential, reflecting an increased activation of a brain area related to a specific function (Rigoulot et al., 2015). These changes in the brain also reflect differences in behavior between musicians and non-musicians. For example, modifications in the temporal lobe (which includes the auditory cortex), can correspond to improved sound perception capacities, and increased grey matter in the frontal areas can be linked to executive functions.

A review by Pantev & Herholz (2011) exploring brain plasticity in the auditory cortex, underlined that the type of tone used in cross-sectional studies provokes different changes in the brain. Indeed, we observe not only a difference in neural activation between musicians and non-musicians (Besson et al., 2007; Pantev et al., 1998; Fujioka et al., 2004; van Zuijen et al., 2004;

Herholz et al., 2009) but also (musical) timbre-specific enhancement for musician participants. In fact, in studies that compared different types of sound input, for example, sine-tones and piano tones (Pantev et al., 1998), neural activation in musicians was significantly different between these two stimuli types. The effect of timbre specificity is thus quite strong, and changes in cortical representations for tones of one's own instrument differ from those of a different instrument (Pantev et al., 2001; Margulis et al., 2009). This is shown not only in the auditory cortex but also through fMRI studies (Margulis et al., 2009), in electrical responses (Shahin et al., 2008), and in the brainstem (Strait et al., 2011). Tests for these timbre-specific modifications, which generally focus on the piano, the flute, and the violin, are also replicated for participants who participated in monitored "short-term" training for both piano (Shahin et al., 2008) and violin (Fujioka et al., 2006). This specificity is not exclusive to single tones, as it has been shown to apply to longer melodic segments as well, including pitch contours, intervals, and polyphonic melodies (Fujioka et al., 2006). Herholz & Zatorre (2012) develop the idea of instrument specificity further in their review, underlining instrument-specific modifications in terms of both auditory stimulation and motor functions. Indeed, in Lappe et al. (2008; 2011) there is a notable difference in MEG activation between groups having studied piano for 2 weeks (auditory and sensory training) and those having only auditory training (listening and commenting on the recordings of the other group), reflecting an instrument-specific effect of musical training on the motor network. These changes can be seen in terms of the size and shape of the hand representation of the motor cortex as well as the lateralization, which differ between pianists and violinists (Bangert & Schlaug, 2006) and extends to other instruments, including voice (Kleber et al., 2009). Such differences can even be stimulated through audio-visual representations (Proverbio & Orlandi, 2016).

Thus, even after short-term training, musical experience provokes anatomical modifications as well as increased activation in the brain. However, these prove to be specific to instrumental training, corresponding to instrument-specific representations in the motor cortex and processing of musical timbres.

1.3.2.2 The influence of musical training on speech perception

Thus, as we have suggested, modifications present in the brain due to musical experience also impact behavior, as shown in various behavioral tests.

Phonemes

The effect of musical experience on phonological perception has been observed for both children and adults. In children, studies have shown a correlation between reading skills, phonological awareness, and musical skills (Anvari et al., 2002, Bhide et al., 2013, Moreno et al., 2009, Skubic et al., 2021). Gordon et al. (2015), who performed a meta-analysis on the transfers between music and language with a focus on literacy skills in children, shows the development of phonological awareness through music (though the effect of these musical skills have a significantly different impact according to the intensity of musical training, Eccles et al., 2020). In adults, music is also shown to impact phonological awareness, where adults with deficits in musical pitch recognition (tune-deaf or tone-deaf) show decreased phonological and phonemic awareness (Jones et al., 2009). The impact of musical training on phonological processing has also been tested on foreign speech, where several studies show a correlation between phonological proficiency in foreign speech and musical training (Slevc & Miyake, 2006). This applies to both phonological pronunciation (Milovanov et al., 2010) and supra-segmental discrimination (Sadakta & Sekiyama,

2011), and extends to other elements of foreign speech such as changes in pitch (Marques et al., 2007).

Tone languages

Studies on tone-language perception also show a number of musical advantages. Generally, findings indicate that non-tone language speakers with musical experience identify tones in tonal languages better than participants without musical experience (Gottfried & Riester, 2000; Gottfried et al., 2004; Alexander et al., 2005; Lee & Hung, 2008; Bidelman et al., 2013; Han et al., 2019). This advantage is robust, as musicians' recognition capacity outperforms non-musicians even when the F0, the frequency on which Mandarin tones are based, has been removed, forcing musicians to reconstruct the fundamental frequency according to the harmonics within the signal (Lee & Hung, 2008). This advantage is equally apparent in words, where musicians recognized both tonal and segmental differences in 4-word sequences better than non-musicians (Marie et al., 2011). Performances by musicians who speak a non-tone language have even been compared to those of native tone language speakers (Alexander et al., 2005; Hutka et al., 2015; Bidelman et al., 2013).

Explanations for these advantages include processing tones as melodic or linguistic information (Delogu et al., 2006), enhanced F0 recognition (Lee & Hung, 2008), or improved learning effects for different types of tones (Ong et al., 2017). Some also attribute these musical advantages to the following: better timbre discrimination, suggested by enhanced MMN (mismatched negativity) responses found for music and speech (Hutka et al., 2015, Martínez-Montes et al., 2013), more robust pitch tracking in brainstem responses compared to both non-musicians and Mandarin speakers (Bidelman et al., 2011); or to auditory enhancements (such as pitch acuity) and better memory (see Bidelman et al., 2013). Intartaglia et al. (2017), whose findings show that musicians'

neural encoding of acoustic information is like that of native tone-language speakers, support this final hypothesis.

Modified Speech perception

When considering modified speech, neural encoding for musicians has generally proved to be more resistant compared to non-musicians in difficult listening conditions. This is the case for speech tested in different reverberation conditions (Bidelman & Krishnan, 2010) as well as for speech in noise. In such conditions, a capacity for more fine-grained frequency discrimination (such as the F2 and F3 contour; Varnet et al., 2015) and better working memory (Parbery-Clark et al., 2009a) allowed musicians to perform better than non-musicians. A more robust and accurate representation of target acoustics and stimulus harmonics (timbre), as well as an earlier response onset timing and phase locking due to enhanced subcortical representation of speech sounds, has been shown to help with parsing melodies from background sounds (Parbery-Clark et al., 2009b). Strait & Kraus (2011) suggest that musician advantages for deciphering speech in noise may be a reflection of improved attention skills, due to strengthened brain networks for selective auditory training. Because of these skills, musicians have often been compared to bilingual participants (D'Souza et al., 2018). Finally, in Meyer (2015), a few participants with musical experience were included in a whistled vowel perception test, where they showed a slight advantage over other participants.

1.3.3 Musical experience and the musician

These numerous studies showing an advantage for participants with musical experience often compare “musician” groups, or participants with “more” musical experience over non-musicians (Eccles et al., 2020, Jakobson et al., 2008; Ho et al., 2003). However, these groups are poorly defined, not only because of a lack of homogeneity in skill sets among “musicians” from various experiments but also because the definition of the musician is in itself complex. In this section, we explore what it means to be a musician, how this term is used in experimental studies, and what specific skills are involved in musical training.

1.3.3.1 Defining a musician

What is a musician? According to the Cambridge Dictionary, a musician is “someone who is skilled in playing music, usually as a job”. Though this definition varies slightly according to the source, the reoccurring qualification for a musician is the skill level obtained which allows one to play increasingly difficult music. The importance of this skill set is reflected in the teacher’s expectations. We can cite a music teacher interviewed by Mills (2010, p.47) who emphasizes the technical skills involved in being a musician. “To me, a musician is someone who can play all the scales on whatever instrument they play, and knows all the key signatures, that are able to play inversions, who show that they are musicians, not that they love music”. This definition not only underlines the acquisition of specific skills to illustrate their “musician” identity, but it also differentiates passion for music and musical skill. In “Musical Identities” (Gracyk, 2003), this differentiation is highlighted according to one’s relationship with music. The first relationship, “identities in music”, reiterates the definition of the “musician” proposed previously, associated

with a long-term practice and manifested through one's identification with their musical instrument. In this definition, the "musician", whose identity is music, is distinguished from the "non-musician", where music makes up your identity. However, Gracyk (2003) criticizes this "musician threshold", underlining that musical identity may be more complex than the one defined by the amount of theorized skill, especially when considering musicality.

The idea of musicality (or having a certain sensitivity to the music), appears in the works of Hargreaves et al. (2002; 2011), who extend the definition of the musician by integrating the concept of "musicianship", the socially and culturally defined concept of a musician. According to Hargreaves et al., the definition of a "musician" extends further than a skill set and often requires participation in certain activities as well as specific social behavior. These activities would include time spent playing regularly in bands, or rehearsing regularly, i.e. participating in the musical world. Thus, being a "musician" is a socially and culturally defined concept and identity that requires skill and expects one to behave and participate in a certain way. Zhang et al. (2020), who considered the musician within the context of psychological experiment, described the musician according to the Three Component Model of the Musician Definition, the three components being skill, identity, and predisposition. Therefore, throughout these definitions the role of skill remains constant even though a number of other factors contribute to the "musician" label. This suggests that the skill set must be defined in order to distinguish musicians from non-musicians.

The Musician as an experimental subject

Most experimental research qualifies their "musician" participants differently, notably because the skillset required is not well established. Indeed, as described by Smit et al. (2023), the criteria that define the opposition between "musicians" and "non-musicians" is relatively arbitrary,

often based on the number of years of musical experience (which is not always reflective of skill). Zhang et al. (2020) proposed a review of publications specializing in music psychology from 2011 to 2017 to compare “musicians” and “non-musicians” and their evaluation methods. They underline a variety of labels used in these different studies, which reflect various types of musical expertise, and therefore of musical evaluation. We can regroup some of these qualifications according to three common evaluation methods (see Annex, A.1, Table 1).

The first of these methods (1) determines musical experience or musical capacities using self-evaluation: participants are requested to describe their level of expertise and their musical capacities. When using a self-evaluation method, authors often define certain thresholds, for example, the minimum number of years of instrumental practice or a minimum age. The precise number of years of experience required can vary significantly between experiments, ranging between 2 and 12 years, though often set around 6 years (Smit et al., 2023; Zhang et al., 2020). This also applies to the minimum starting age, which can be anywhere between 7 and 13 years old. Other factors, such as currently playing the instrument can also be included. The second method (2) used for evaluating instrumental skills, is having a degree or diploma obtained in a musical institution[1]. Finally, the third method (3), uses certain tests for defining musical levels including the AMMA (or Advanced Measures of Musicality Audiation), Wing test, or the Goldsmith Musicality Index. These tests have different primary objectives. The AMMA, created by E. Gordon (1989), seeks to measure musical aptitude using “audiation”, which tests tonal and rhythmic skills. The “Wing” or “Wing Musical Aptitude Test”, was originally used to find musically bright children (Wing, 1962). Finally,

[1] This method is easily paired with recruitment for behavioral experiments, as participants from musical institutions will have completed an audition to enter and will complete an exam to graduate.

the Goldsmiths Musical Sophistication Index, made to adapt to experiments including musicians, is a psychological survey (therefore self-evaluating) that also includes an auditory test. This index allows users to factor in the criteria according to the needs of the experiment.

There are some advantages and disadvantages to each of the methods proposed. By using the self-evaluation method (1), participants can freely disclose all types of musical experience, without limiting musical experience to a specific style or form. This also reflects how individuals consider their musical level, a declaration of one's own musical identity. The disadvantage of this method is that it can lack a true measure of one's instrumental and auditory skills, as participants declare their musical skills subjectively (Smit et al., 2023). By requiring a minimum age or number of years of musical experience, this creates a clearer and more quantifiable measure of musical knowledge and conforms to part of the definition of "musical identity". However not only do the age limits and number of years of experience vary significantly according to the experiment, but the number of years spent playing and the age at which one begins music are not always proof of skill. Indeed, 6 years of musical experience seems quite minimal (Zhang et al., 2020), especially if there are no explicit learning conditions surrounding these years of musical instruction. This dilemma is resolved by asking for a musical diploma or enrollment in a music program (2), which obliges one to have obtained a certain quantity of musical skills, often also necessitating an early start to one's musical education. There remain, however, slight differences between different music schools, and types of musical education obtained. In addition, though the institutionalized qualifications for music are generally associated with Western classical music[1], other schools may provide diplomas

[1] In most universities around the world, western music performance qualifies as a degree obtained in specialized music schools or general universities, who administer diplomas following the Bachelor's, Master's, and Doctorate system..

in non-classical music or non-western music (though some non-western cultures do not highlight such a large distinction between musicians and non-musicians). Finally, musical tests (3) can allow for an objective measure of musical capacities, similar to musical diplomas. However, the tests function differently, as described above, and are often orientated towards classical music expertise (see Verdis & Sotiriou, 2017 for a study on the applicability of AMMA on non-western music), and do not consider the type of musical experience and instrument specialization. Finally, a musician's skill is not entirely based on rhythm and pitch recognition, or musical analysis results, as the product of musical learning is performance, which is not a part of such a test. Indeed, the biggest downfall of musical tests is that the skills tested may not correspond to the skills in the task, and they are not fully representative of musical expertise, especially concerning instrument skills. In addition, though the musical instrument played is sometimes a factor included in descriptive participant data, most musicians are grouped together despite having different instrument specializations (see Annex, A.1, Table 1).

1.3.3.2 Musical skills

Thus, to understand the "musician advantage" in behavioral studies, we can consider one's musical (and instrumental) skills more closely. In France, classical music is taught in a highly institutionalized conservatoire system, giving a relatively clear understanding of classical music skill sets. Classical music training often starts very early, as encouraged by certain pedagogies (About the Suzuki Method, 2022) and implemented by music schools, who sometimes impose a maximum age limit for starting an instrument in music schools (this is the case for the music school in Nice, France

for example[1], Informations Générales, 2022). Students can start introductory classes at music schools in France as early as 5-6 years old, where they learn about musical instruments, sounds and vocabulary, group singing, artistic expression, and movement. This introductory course facilitates entry into the 1st of 3 cycles, generally at age 6 or 7. These cycles last 3-5 years and include a yearly examination, ending with an overall evaluation which will allow students to enter the next cycle. The amount of time students are required to spend studying music increases per cycle, starting with 2-5 hours of music classes a week in cycle 1, increasing to 4-7 hours for cycle 2 and again in cycle 3. Additional classes include ear training, music history/culture, and singing/instrumental groups. After the 3rd cycle, students choose to either become high-level amateur musicians, with or without a CEM (Certificat d'Etudes Musicales) diploma, or to pursue music as a profession. To do so, one must complete 2-4 additional years of musical training and receive the DEM (Diplôme d'Etudes Musicales)[2] concluding their studies in music schools (conservatories). Once the DEM is obtained, students can then continue into higher-level professional musical education (Bulletin, 2006; Tableau des cycles d'études en conservatoire, 2018). Such courses will usually include longer individual lessons, chamber music, ear training, musical analysis, orchestra, and history of music. In France, two different types of professional music schools exist (the National Conservatories and the Pôles Supérieurs), both of which require an entrance audition. These schools allow students to obtain a professional diploma equivalent to a Bachelor's degree in Music Performance in 3 years (1 350 hours) (Métiers de la Musique, 2018), which can then be followed by a Master's degree or even a Doctorate. Musicians who play other musical styles or genres do not always follow the same

[1] This music conservatory requires beginner musicians to be within the CE1 – CM1 age bracket (CRR Nice), where children are usually between 6 and 9 years old.
[2] This diploma is now called the "DNEM" (Diplôme National d'Etudes Musicales), though we will use its previous title as it is present throughout the thesis (Tableau des cycles d'études en conservatoire, 2023)

academic path as classical musicians, as music schools in France only started teaching Jazz and Popular Music in the 1980's-1990's.

The musical skills acquired can therefore be divided into two groups: instrumental performance classes and non-performance classes (theory, musical analysis, history, composition...). Instrument specialization and performance start at the beginning of the 1st cycle and take up the most amount of time in the student's schedule (including individual lessons, group ensembles, and practice time). These skills are tested at the end of each cycle with an instrument-based performance: a 20-minute exam including a performance portion and a sight-reading exercise in Cycle 1, a 45-minute exam including a 6-10 minute performance, a sight-reading task, a prepared singing task and a musical analysis task in Cycle 2, and a 20-minute instrumental performance in Cycle 3. These exams test both instrument skills (which includes instrumental techniques and personal development linked to the instrument) and general musical skills (playing the correct notes/rhythm, underlining musical phrasing, mastering different musical codes, showing knowledge of musical culture and musical structure). In lower levels, these skills are separate exams, however, starting in cycle 3, the musical performance tends to be judged in its entirety, as, in a performance, such complementary skills are almost inseparable. Thus, the skill sets of each level of musical skill correspond to a different form of knowledge.

> Due to the similarities between music and speech (notably in terms of pitch and timbre), a transfer between musical experience and speech can be established. Such transfers first show effects on parts of the brain, creating a musical "advantage" for parts of speech perception, including phonological awareness and modified speech perception. It is,

however, difficult to understand which elements in music allow for such an advantage, as well as what specific components of speech perception are affected. For these reasons, we sought to further define the skills of classical musicians in France, by rethinking the way musicians and musical skills are measured and by including instrument specialization. This will allow us to explore elements such as instrument timbre or production methods in speech perception.

1.4 Recapitulation and Questions

1.4.1 Recapitulation

In this first chapter, we have explored three main themes that we will use to construct the experiments proposed in this thesis: speech perception, whistled speech, and musical experience.

We first focused on the role of the phoneme (a controversial pre-lexical unit), used to form a relationship with higher processing levels and smaller units. After exploring the phoneme's multimodal role and comparing it to various other units, we then proposed a more detailed characterization of the phoneme in terms of acoustic and articulatory elements, which differentiate vowels and consonants (Chpt.1.1.1). Finally, we sought to deconstruct the phoneme and consider the role of various speech cues that it represents. One approach is to recreate the speech signal with as few cues as possible (synthetic speech). This reduction of speech gave way to various theoretical perspectives on the role of articulatory and acoustic cues. We then suggest that by using forms of modified speech, we can test speech perception cues. We propose using whistled speech, as it reduces the acoustic speech signal, all while maintaining articulation, and natural variation between speakers (Chpt.1.1.2).

Whistled speech, a natural form of modified speech used in isolated and geographically extreme regions to communicate over long distances, has been documented around the world. Whistled speech augments spoken modal vowels to higher whistled pitch ranges, which vary according to the speaker's voice, and attributes a certain range of frequencies to each vowel. The whistled consonants maintain spoken articulation and modulate these relatively fixed vowel

frequencies according to the articulatory movements of production (Chpt.1.2.1). We took an interest in production techniques and phonological groups of the non-tonal whistled speech form used in the Canary Islands (known as Silbo), as Silbo is either used directly in the experiments proposed, or the same whistled transposition is applied to French. Various perceptive tests have previously been conducted on whistled speech, and these serve as the starting point for the behavioral tests included in this thesis. They also help to direct the way in which whistled speech can be used as a tool to study speech perception (Chpt.1.2.2).

Finally, we consider the effect of listener variability by taking a specific interest in musical experience and exploring how musical skills can be transferred toward speech perception. Previous studies have been robust in showing that musical experience transfers skills towards very similar tasks (Chpt.1.3.1). Thus, for there to be an effect of musical experience on speech perception, the two processes must be similar enough to transfer skills from one domain to another. Indeed, the similarities between speech and music start with the signal itself, which comprises many similar components, including pitch, timbre, and melody. These similarities justify a transfer between music and speech, as seen through changes in the brain, and as a consequence, advantages in speech perception (Chpt.1.3.2). However, to understand the transfer that takes place between speech and music, it is essential to define the "musician" as a label, including how one acquires such a label, and what skills are learned (Chpt.1.3.3). These explorations also allow for a more targeted measure of musical experience, providing a deeper understanding of its effect on whistled speech perception.

1.4.2 Questions

By taking an interest in these two forms of variability in speech perception, both within the speech signal transmitted and in the experience of the receiver, I sought to answer several questions, exploring both the process of whistled speech perception and the effect of musical experience: (1) Are naive listeners able to use their knowledge to correctly categorize whistled speech at different perceptual levels (phonemes and words)? (2) What cues are used in this categorization process? (3) Will musical experience have an impact on whistled speech perception? (4) Which elements of the speech signal help provide better results for this group of participants? (5) How do these reflect specific skills acquired through musical training such as musical level achieved and instrument specialization?

These questions will serve as a guideline through the various articles and experiments presented in the following chapters.

1.4.3 Article description

In Chapters 2 and 3, we explore whistled speech perception by naive listeners (listeners who had never heard whistled speech previously), by focusing on the whistled phoneme.

We first explore the whistled consonant, which has never been tested previously with naive listeners (Chpt.2), by focusing on the consonants /k/, /p/, /s/, and /t/ presented in the form of VCV nonsense syllables. The first article proposed here (Chpt.2.1 -Tran Ngoc et al., 2020a) considers whether naive listeners (participants who have never heard whistled speech before), can categorize consonants correctly. We also include certain aspects of intra-whistler variability, the possibility of a learning effect and question the importance of the whistled pitches themselves. The second article

of this chapter further develops upon naive listener categorization by considering consonant confusion rates (Chpt.2.2 - Tran Ngoc et al., 2022a), thus allowing for a more precise comparison between the consonants chosen.

In Chapter 3, we explore the whistled vowel, which has been tested previously in several different experiments (see Meyer, 2008; Meyer et al., 2017). The analyses are proposed here on two different perceptual levels (the phoneme and the word), targeting the same whistled vowels as previous studies (/i/, /e/, /a/ and /o/) and including naive listeners. The first whistled vowel experiment proposed here (Chpt.3.1. - Tran Ngoc et al., 2020b) seeks to replicate previous findings and to explore the effect of inter-whistler variation by including two different whistlers, thus considering the possibility of a learning effect between whistler productions. It is followed by a supplementary study, where we compared the learning effect according to each whistler (Chpt.3.2 – Tran Ngoc et al., 2023e). Finally, we explored the role of these same target whistled vowels within the context of the whistled word (Chpt.3.3 - Tran Ngoc et al., 2023a).

In Chapters 4 and 5, the focus turns towards the effect of musical experience on whistled speech perception, elaborating upon the previously described articles and experiments. This exploration considers both the effect of perceptual variability of the listener, all while furthering our understanding of the skills gained through musical experience and how these skills may transfer towards speech perception.

In Chapter 4, we first consider the effect of musical experience on the perception of whistled phonemes, reprising the themes of Chapters 2 and 3. The first article of this chapter (Chpt.4.1 - Tran Ngoc et al., 2023b) considers the effect of musical experience on isolated whistled vowels, taking time to consider how whistler variability and the various pitches of the vowels affect the

performance of musician participants. We also wonder if whistled speech - whose cues may resemble music at first glance- is treated as music or as speech by musician participants. We then considered the effect of musical experience on whistled consonant categorization (Chpt.4.2 - Tran Ngoc et al., 2023c), exploring the transfer that occurs between whistled speech and music by opposing a general skill transfer and a sound-specific auditory transfer (as represented by musical instrument specialization).

Finally, in Chapter 5, we consider the effect of musical experience on whistled word perception (Chpt.5.1 - Tran Ngoc et al., 2023d), further developing upon the consonant and vowel correspondences. To do so, we compare high-level musical experts with expert Silbo whistlers, differentiating musician participants according to instrument specialization, and analyzing the effect of instrument-specific skills in the context of the whistled word recognition. For an overview of the articles included in each chapter, see Table 3.

Table 3:

Overview of articles and experiments in each chapter

Chapter	Article name	Experiment	Target theme
2.1	Tran Ngoc et al., 2020a	Expt 1A	Whistled consonants
		Expt 2	Low whistled consonants
2.2	Tran Ngoc et al., 2022a	Expt 1B	Whistled cons confusions
3.1	Tran Ngoc et al., 2020b	Expt 3	Whistled vowels (2 whist)
3.2	Tran Ngoc et al., 2023e	Expt 4	Whistled vowels (2 whist)
3.3	Tran Ngoc et al., 2023a	Expt 5	Whistled words
4.1	Tran Ngoc et al., 2023b	Expt 6	Musical experience and vows
4.2	Tran Ngoc et al., 2023c	Expt 7	Musical experience and cons
5.1	Tran Ngoc et al., 2023d	Expt 8	Musical experience and words

Chapter 2

Whistled consonant perception

Introduction

The first theme of this thesis, whistled speech perception by naive listeners, is addressed in Chapters 2 and 3, where we have regrouped articles according to their focus on either consonants or vowels. Despite this separation, we believe that these chapters can nevertheless be regrouped around the theme of phonological perception. We question whether naive listeners can correctly categorize and recognize whistled phonemes and words, providing insight into the specific whistled cues used and the effect of variability in whistled speech perception.

In this chapter (Chpt.2), we explore naive listeners' ability for whistled consonant categorization in two articles. As whistled consonant categorization has never been tested before with naive listeners, and given that previous experiments with whistlers (see review Chpt.1.2.2.2) showed several inconsistencies (including different stimuli, unknown distances, etc), we tried to provide a clear and reproducible experimental context. In these two articles, we consider the whistled consonant in a vowel-consonant-vowel form (VCV), with the same vowel heard before and after each consonant. This format was previously used by Rialland (2005), though with varying vowel contexts, thus already providing an idea of the consonant cues available to listeners. Here we explore the cues of a single whistler, by taking an interest in the construction of these consonant

groups and how they are used by naive listeners. We address several specific objectives and questions in these two articles.

After first considering whether consonants can be categorized correctly by naïve listeners, we then wonder how participants will treat the different characteristics and cues of each whistled consonant targeted here. Can they be distinguished from each other? And if so, what element is used to distinguish them? Finally, how are naive listeners impacted by intra-whistler variability, and can they learn to improve consonant categorization through training?

These themes are treated in both articles, which contain many similarities, including identical experimental designs (3 parts, including a 2nd part with feedback and a 3rd part with intra-whistler variability). Each article, however, treats consonant categorization differently. The first article (Tran Ngoc et al., 2020a) considers whistled consonant categorization by questioning the role of the whistled pitch more specifically. This article contains two experiments. The first experiment (Expt 1A) includes listeners who participated online (using PCIbex Farm) by listening to natural unmodified VCV consonant stimuli. The second experiment (Expt 2) considers whistled consonant stimuli that have been lowered to a speech-like pitch. Participants from this second experiment, completed this experiment in person through PsychoPy.

In the second article (Tran Ngoc et al., 2022a), the focus turns toward the listener's whistled consonant confusions. Previously applied to whistled vowels (see Meyer et al., 2017), this perspective provides insight into the phonological boundaries between consonants defined by acoustic cues. The experiment included in this article (Expt 1B), is heavily based upon the previous article's first experiment (Expt 1A), as it is also conducted online and uses an identical design.

However, this experiment (Expt 1B) included 10 more participants than experiment 1A, as presented in Table 4.

Table 4:

Description of the experiments in articles 1 and 2 of Chapter 2

Chapter 2 - Consonants	Article name	Experiment	Target	Design	Partici-pants	Location
2.1	Tran Ngoc et al., 2020a	Expt 1A	Whistled consonants	3 parts	20	Online
		Expt 2	Low whistled consonants	3 parts	16	In person
2.2	Tran Ngoc et al., 2022a	Expt 1B	Whistled cons confusions	3 parts	30	Online

2.1 Categorization of whistled consonants by naive French speakers

Abstract

Whistled speech is a form of modified speech where some frequencies of vowels and consonants are augmented and transposed to whistling, modifying the timbre and the construction of each phoneme. These transformations cause only some elements of the signal to be intelligible for naive listeners, which, according to previous studies, includes vowel recognition. Here, we analyze naive listeners' capacities for whistled consonant categorization for four consonants: /p/, /k/, /t/ and /s/ by presenting the findings of two behavioral experiments. Though both experiments measure whistled consonant categorization, we used modified frequencies -lowered with a phase vocoder- of the whistled stimuli in the second experiment to better identify the relative nature of pitch cues employed in this process. Results show that participants obtained approximately 50% of correct responses (when chance is at 25%). These findings show specific consonant preferences for "s" and "t" over "k" and "p", specifically when stimuli is unmodified. Previous research on whistled consonants systems has often opposed "s" and "t" to "k" and "p", due to their strong pitch modulations. The preference for these two consonants underlines the importance of these cues in phoneme processing.

Article Information

Article Status

This article has been published in the proceedings of INTERSPEECH 2020:

Tran Ngoc, A., Meyer, J. & Meunier, F. (2020a). Categorization of Whistled Consonants by Naive French Speakers, *INTERSPEECH 2020 – 21th Annual Conference of the International Speech Communication Association, September 14-18, Shanghai, China, Proceedings*, 1600-1604. https://doi.org/10.21437/Interspeech.2020-2683

Keywords: consonant categorization, whistled speech, whistled languages

Categorization of Whistled Consonants by Naive French Speakers

Introduction

Whistled speech is a naturally modified speech form characterized by its frequency augmentation and a whistled transposition of certain features encoded in the modal speech spectrum, drastically changing the spoken timbre. Whistled vowels of non-tonal languages often employ generally stable frequencies, which depend on the whistling technique, the language, the whistler, and the vowel position (Meyer, 2005; Meyer, 2015). The consonants modify these vowel frequencies, adding stops and pitch changes as the whistlers "pronounce" the consonants while whistling. We can consider whistled speech akin to other forms of modified speech, where naive listeners are able to identify and categorize certain aspects, such as phonemes (Blanco et al., 2018).

Whistled speech recognition and categorization experiments first started in the 1960-70's on Bearnese and Turkish, however naive listeners were not tested and these studies focused on words or logatomes [Meyer, 2015; Busnel & Classe, 1976). In 2005, Rialland ran a behavioral experiment on VCV logatomes whistled and identified by Spanish whistlers while standing 15m apart, obtaining 57% of correct answers with better responses for certain consonants and vowels (Rialland, 2005). More recently, Meyer et al. (2019) conducted a syllable recognition experiment (/ta/, /da/, /ka/, /ga/) with Tashlhiyt Berber whistlers to test the dental-velar contrast and evaluate the impact of the absence of voicing on whistled consonant recognition. Tests on naive listeners only date back to 2005. Such studies included participants of different language backgrounds (Spanish, French, Chinese) and a whistled vowel recognition paradigm based on

Spanish vowels, obtaining results well over chance for all categories of listeners with striking differences between language background and vowel positions (Busnel & Classe, 1976; Meyer et al., 2017). This success causes us to question whether this naive listener capacity for recognition and categorization also applies to whistled consonants. We thus tested naive French speakers' categorization capacities for whistled Spanish consonants through two behavioral experiments. This also allowed us to explore other complementary questions: can naive listeners learn to categorize whistled consonants? Which factors or methods underlie participants' consonant categorization?

To answer these questions, our experiments contain three parts: the first part asks participants to categorize the whistled consonant stimuli without any feedback or presentation, the second presents the whistled consonants and provides feedback, and the 3rd part follows a similar structure as the first part, but includes several natural variations of each consonant using different recordings. This allows us to test whether participants learn to apply consonant models to multiple varieties of each consonant, a method suggested by the results of Hervais-Adelman et al. (2008), where perceptual learning generalized to untrained word stimuli is observed for noise-vocoded speech. To understand the mechanisms for consonant categorization, we will compare previously suggested whistled consonant systems with the participants' responses.

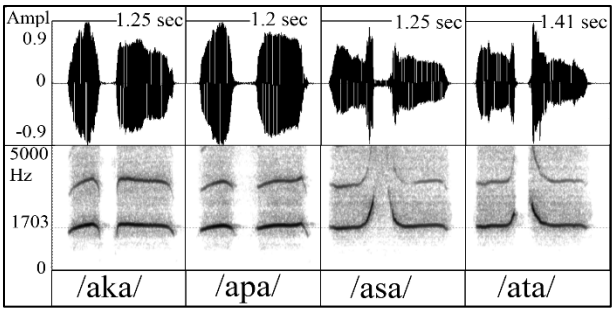
The whistled consonants chosen (/p/, /k/, /s/ and /t/[7]) and recorded in Silbo (the whistled Spanish of the Canary Islands), have often been grouped together based on their articulatory loci, as well as frequency and/or amplitude modulations. Trujillo, for example, proposed 4 consonant

[7] The choice of representation with "//" previously used by Meyer, 2015, suggests a phonological representation of the whistled phoneme

groups and Rialland 8 groups, both opposing whistled /p, k/ to /t, s/ either regrouping the manner of whistling (Trujillo) or consonant perception (Rialland argued for higher loci in /s/ than /t/) (Rialland, 2005). It is important to note that these supposed groups derived from observed phonetic reductions are partly dependent on the whistling technique, the position of the consonant in the word, the speech rate, and the proficiency of the whistler (Meyer, 2015), parameters that previous studies did not systematically control. However, all researchers agree on two clear distinctions among whistled consonants in non-tonal languages: one between consonants with high (/s/ and /t/) or low (/p/ and /k/) whistled loci, and one between continuous (/s/) and non-continuous consonants (/k, p, t/). High loci systematically correspond to consonants rising after the previous vowel (V1) and falling towards the next vowel (V2) (see /asa/ in Figure 1), and low loci the reverse (Meyer, 2015). The classification of /s/ is more complex because it emulates the continuous fricative aspect of spoken speech, which is expressed by a low amplitude continuation of the whistled sound. Thus, whistled fricatives can be considered as non- or semi-continuous, depending on the speech rate (in faster speech fricatives seem continuous because of their more gradual amplitude envelope modulation (Meyer, 2015) and the listening distance.

Figure 1:

Spectrogram and signal of VCV forms



When considering the directives given to students learning the Silbo language, those of La Gomera Island follow recommendations based on Trujillo's groupings, whereas those of Yo Silbo association, the most active Silbo revitalization association in the Canary Islands, assemble the consonants into five pronunciation-based groups using VCV configurations. This classification opposes /t, s/ to /p/ and to /k/ (Diaz, 2008) which may take into account the glottal occlusion that can be heard more easily in /k/ than /p/, or the bilabial attack after the consonant stop in /p/. The clarity of the stop could also be a defining commonality in /t/ and /k/, which is not present in /s/ and /p/. The occlusive and constrictive consonant opposition is not proposed as a main cue, but certain very skilled whistlers manage to develop it (Diaz, 2009), thus, it is considered as a secondarily developed opposition. These models also allow us to justify our choice in consonants and oppose these consonant cues, which could be key to establishing categorization methods.

The second experiment follows the same structure as the first, using modified consonant frequencies in an effort to pinpoint the importance of these categorization cues in spite of a drastic frequency shift. Though these experiments target whistled speech, the natural modification of speech cues reflects more generalized phoneme processing methods as well as subconsciously defined phoneme categories.

Two groups of participants performed the whistled phoneme (consonant) identification tasks, the first with natural whistled consonants (Experiment 1), and the second with modified stimuli (Experiment 2).

Experiment 1

Method

Stimuli

We chose to test four distinct consonants of spoken Spanish, that have either identical or easily learned pronunciation differences in Spanish and French (Molina-Mejia, 2007): three occlusive consonants ([p]-bilabial), [t]-dental/alvéolar), [k]-velar) and a fricative ([s]-alveolar), followed and preceeded by the vowel [a], giving the following V1CV2 forms where V1=V2=[a] : [ata], [aka], [asa] and [apa]. The use of a VCV form is justified as it reduces variations due to Consonant and Vowel co-articulations at play in whistled Spanish (Meyer, 2015). Four instances of each of the four /aCa/ segments were whistled by the same proficient whistler-teacher of Silbo (the whistled Spanish of the Canary Islands) and recorded by Julien Meyer[8].

In experiment 1, the frequencies before and after each consonant closure vary between 1141.9 and 2628.7 Hz, with an average of 1715.86 Hz. These frequencies usually reflect the frequency shapes of the 2nd and/or 3rd speech formants, though not necessarily their frequency values, due to a different sound production process (such as a more closed mouth) (Meyer, 2015).

Procedure and Design

Experiment 1 was programmed using PCIbex Farm and took place online from participants' own homes. Before starting the experiment, participants were asked their age, the languages they speak (and their level), as well as if they play any musical instruments. As Experiment 1 was online,

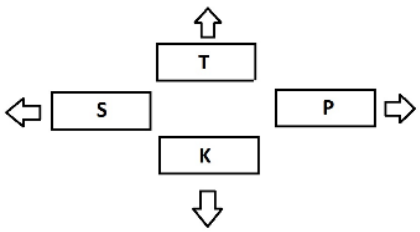
[8] See Annex, A.2, Table 2 for the each stimuli used in Experiment 1, presented with the acoustic signal and the spectrogram.

they were to indicate whether they used headphones, earbuds or speakers, to give the name of the brand and were to adjust the volume to a comfortable listening level. We recruited the participants through various social media networks, considering, once we excluded self-declared speech/hearing impairments, that participants did not have any pre-disposed differences in performance.

During part 1 of the experiment, participants first listen to an example of whistled speech to introduce them to the acoustic specificities of whistled signals. The four /aCa/ recordings presented (one of each consonant, see Figure 1) are used during part 1 without any indication of the consonant heard. These four recordings were chosen according to the stability of whistled vowel frequencies surrounding the consonant. The participants then hear these clips in a random order and are asked to respond with either “p”, “k”, “t” or “s” after each clip. These consonants are attributed to the arrow keys on the keyboard according to the layout of both azerty and qwerty keyboards. Participants see Figure 2 on screen as they listen and respond to the 40 recordings (10 times each consonant) which make up part 1.

Figure 2:

Consonant/Arrow key attribution



Part 2 is a training phase with feedback, using the same whistled audio tracks as part 1. We first present the four different consonants in a random order by playing a spoken version of the VCV segment, followed by the whistled version. An image of the consonant appears on the

screen simultaneously. Following this, participants complete a shorter version of the previous test albeit with feedback. Participants hear each clip (each consonant) 4 times, amounting to 16 total excerpts. Feedback is given after each response: "*Bravo*" when correct and "*Non ce n'était pas la bonne réponse*" – "No that was not the correct answer", when false.

In part 3 of the experiment, participants hear sound clips and are once again requested to indicate which consonant was heard (using Figure 2). However, in this portion, 3 additional versions of each consonant are included, amounting to 4 total variations per consonant. As this applies to all 4 consonants, 16 recordings are heard, out of which 12 are unfamiliar variations (i.e. not heard in part 1). Each recording is played 3 times and participants hear a total of 48 stimuli in part 3.

Participants

This first study included 20 adults (15 females, 5 males, mean age: 29.0 years, *SD* = 9.78) whose first language was French, who did not have any language or hearing impairments and who did not play any instrument at a high or pre-professional level. Participants gave informed consent before starting the experiment.

Results

Our analysis focused on parts 1 and 3, excluding the short training portion (part 2) due to the small sample size.

We first compared both parts 1 and 3 by taking into account the 40 answers given in part 1 by each participant as well as the 48 answers given in part 3. This gave us 1760 data components

with 51 % of correct answers, i.e. participants categorized the whistled consonants properly. We ran a global repeated measures Anova, that included Consonant type (k,p,s,t) and Part (part 1, part 3) as within fixed variables and participants as a random factor. We observed a significant main effect of Consonant type ($F(3,60)=10.047$; $p < .001$). The main effect of Part and the interaction between the two factors were not significant.

We then ran a post hoc test to look at specific comparisons using a Bonferroni correction in order to perform the multiple comparison test. It appears that "p" is significantly different from "t" and "s" ($p < .001$) and that "k" shows a tendency to be different from "s" and "t" ($p = .1$). This opposes "p" and "k" to "s" and "t" in the following manner: "t" = "s" > "k" = "p".

Discussion Experiment 1

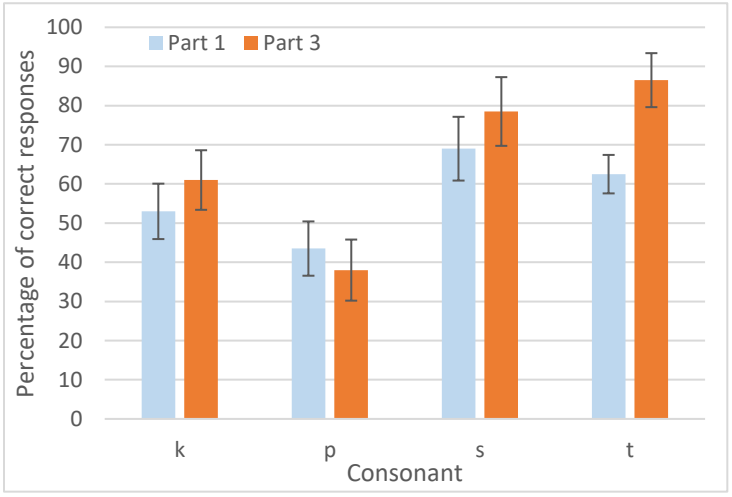
The overall performance shows that participants recognized the set of consonants well over chance. In addition, the hierarchy shows a preference for the consonants with high loci, or those containing a rising pitch towards these loci ("s" and "t", see Figure 1). Considering that parts 1 and 3 were constructed differently, the results from these parts provide insight into the evolution of participants' performances.

We can take a closer look at part 1, which reflects the participants' initial and naive recognition of consonants. Participants succeeded in categorizing the consonants well over chance (46.5% of correct responses for 800 items), however, specific post-hoc comparisons using the Bonferroni correction revealed only one significant difference "s" vs. "p" ($p < .02$). Contrary to the overall hierarchy, it seems that in part 1, two hierarchies could be proposed: "s"= "t" = "k" > "p" or "s" > "t" = "k" = "p".

The lack of difference between Parts in the overall performance, and of interaction between the Parts and Consonant type, could suggest that participants learned consonant categorization, as part 3 included more stimuli variation (with 75% of new stimuli). Though this may be due to other factors, if no learning were to take place, we would expect the results from part 3 to be significantly lower than those of part 1. If we take a closer look at performance in part 3, participants recognized 55% of consonants out of the 960 items. Specific post-hoc comparisons using the Bonferroni correction revealed three significant differences which differ from those of part 1: “p” vs. “s” and “p” vs. “t” (p <.001) and “k” vs. “t” (p <.05). These significant differences suggest a clearer recognition of “t” compared to the other consonants (Figure 3).

Figure 3:

Average correct responses obtained per consonant and participant in parts 1 and 3 of Experiment 1



Experiment 2

In Experiment 2, we used modified frequencies lowered below 600 Hz, a range which is impossible for humans to whistle. This modification is justified by the fact that whistled speech perception, encoded on a simple frequency line, is more “relative” than spoken speech. This bears some

similarities with relative perception in musical instruments, such as the flute, which have simple frequency timbres.

Method

Stimuli and procedure

The stimuli used in Experiment 2 are the same recordings as in Experiment 1, with a modified overall frequency (F/5). These frequencies vary between 228.38 Hz and 525.74 Hz, with an average of 343.17 Hz. This frequency shift was performed using the Gotzen et al. (2000) Phase Vocoder (which also maintains relative amplitude differences but may alter their proportion). While the design and the procedure were the same as those of Experiment 1, we conducted Experiment 2 in person. We tested for the difference between results obtained online and in person in a different experiment, using identical whistled phonemes and stimuli. We found this difference to be negligible (Chpt.3.1 - Tran Ngoc et al., 2020b). All participants heard the stimuli through Sennheiser HD 200 Pro or Sennheiser MB360 headphones and the volume was maintained at the same level for all participants. Experiment 2 was programmed using PsychoPy and took place in a quiet room in the BCL lab (MSHS, Nice, France).

Participants

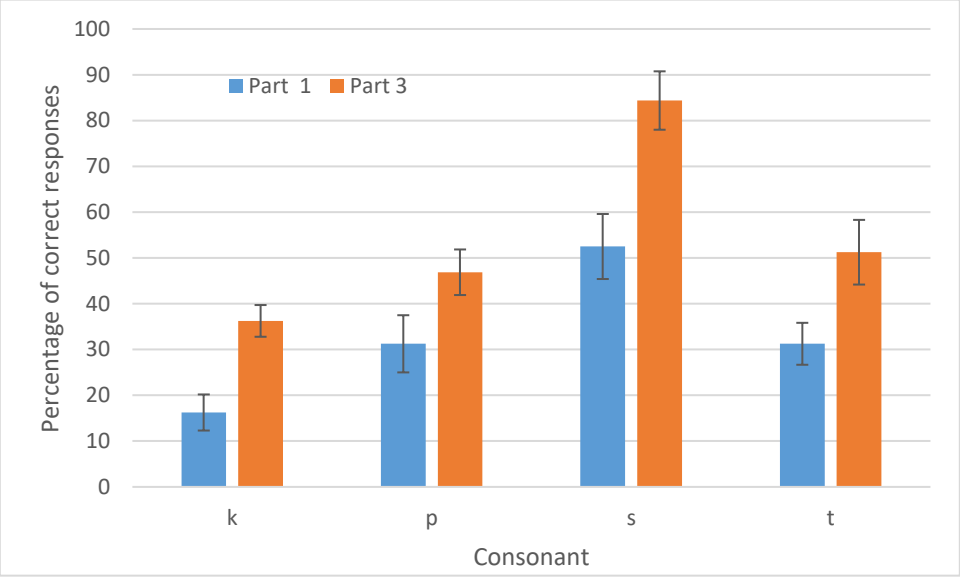
Experiment 2 was completed by 16 participants (9 females, 7 males, mean age: 24.4 years old, *SD* = 5.77) who were native French speakers, did not have language or hearing impairments and were not high-level or pre-professional musicians. These participants were volunteer students recruited from l'Université Côte d'Azur. Participants gave informed consent before starting the experiment.

Results

In our analyses, we took into account the 40 responses given in part 1 and the 48 answers given in part 3 for each participant, amounting to 1408 items. Participants properly categorized the low whistled consonants with 41.5% of correct answers. We first ran a global repeated measures Anova that included Consonant type (k, p, s, t) and Part (part 1, part 3) as within fixed variables and participants as a random factor. We observed significant principal effects of Part ($F(1,15)=6.700$; $p<.05$) and of Consonant type ($F(3,45)=11.409$; $p<.001$). The interaction between the two was not significant. As it can be seen in Figure 4, participants obtained 32.7% of correct responses in part 1 and 45% in part 3. We then ran a post hoc test to look at specific comparisons using a Bonferroni correction ($p<.05$). It appears that "s" is significantly different from "k", "p" and "t", which do not show any significant differences. Therefore, we have "s"> "t"= "k"= "p".

Figure 4:

Average correct responses obtained per consonant in Parts 1 and 3 of Experiment 2



Discussion Experiment 2

The findings above demonstrate that a different consonant hierarchy was obtained in Experiment 2 compared to Experiment 1, underlining a preference for “s” (a high loci continuous consonant). These individual consonant differences are consistent both in parts 1 and 3 of Experiment 2, which, when tested separately, show identical hierarchies. In addition, the greatly improved results of part 3 prove that participants retain models for consonant movement from parts 1 and 2, and apply them to part 3 (especially for “s”).

Comparison Experiment 1 and Experiment 2

Finally, when comparing the results from the two experiments including both data sets in a global Anova with Experiment as a between subject factor, we observed significant main effects of Experiment ($F(1,34)=10.9$, $p<.01$) and Consonant type ($F(3,102)=16.545$, $p<.001$). Two interactions also reach significance: Part*Experiment ($F(1,34)=4.649$, $p<.05$) and Consonant*Experiment ($F(3,102)=5.077$, $p<.01$). Looking at specific comparisons with post-hoc tests, we observed that the amount of correct answers obtained in part 1 is different between the two experiments (46.5% compared to 32.75% and $p<.001$). The significant difference between these experiments in part 1 can be attributed to two consonants: “k” and “t” ($p<.01$). This suggests that a difference in frequency influences the recognition of certain consonant categories.

General Discussion

Overall, whistled consonant recognition averages at 51%, with certain consonants being more difficult to recognize (/p/) and others being easier (/s/ or /t/). The recognition of this modified speech form also applies to lowered whistled frequencies (42% of correct responses for Experiment 2). These results are in line with those obtained by Meyer for vowel recognition (Meyer, 2005), as well as Rialland (2005), where Silbo whistlers showed consonant preferences. In addition, 46.5% of correct responses were obtained for non-modified whistled consonants in part 1 (well over chance, 25%) confirming that naive listeners can categorize the chosen set of whistled consonants. There was no significant difference between parts 1 and 3 in Experiment 1, indicating that recognition rate did not decrease, as it should have if the new stimuli had not been identified. This underlines the fact that participants learn from the consonant model rather than from the recording itself, and that these models can be integrated and applied to more varied forms of elicitations.

Through these experiments, we defined two consonant hierarchies that reflect certain preferences, reprising some aspects of previous research. In Experiment 1 (“t” = “s” > “k” = “p”) the preference for “t” and “s” seems to correspond to the opposition between “high frequency modulated consonants” with high loci (“t” and “s”) and consonants with low loci (“k” and “p”) (Busnel & Classe, 1976). The tendency for “k” to be different from “s” and “t” rather than significant suggests that the clear glottal attack cue, which characterizes “k”, is easier to identify for some. “t” also uses this cue: this may explain the overall facility participants had with the consonant, described by “t”> “k” in part 3 of Experiment 1.

In Experiment 2, “s”> “k”= “t”= “p” (Figure 4) seems to confirm the same predilection for “rising pitch” consonants with high loci or articulation found in Experiment 1, in spite of the change in frequency. Though opposing “s” to “k”, “p” and “t” could underline the identification of occlusive (“s”) and constrictive (“p”/ “t”/ “k”) or the continuous/non-continuous difference, the comparison between both experiments shows a significant difference between “k” and “t”, but not “p”. This suggests that the clear attack cues of “k” and “t” are harder to distinguish in the lower frequencies. This preference for “s” was also present in part 1 of Experiment 1. Does this suggest that continuous sound with pitch change is easiest to identify in extremely modified speech? Or, do participants tend to consider the lowered consonants (which no longer approach the frequency values of the second and third formants) as non-speech sounds, drawing from musical comparisons. Alternatively, is the whistled “s” recognized best because its timbre resembles that of fricatives?

Conclusions

In conclusion, naive French listeners recognize whistled consonants above average and generally use pitch movement to identify the sound heard correctly. This is coherent with the fact that frequency modulations are the most resilient aspects of the signal with better propagation for long distance communication. This capacity may be due to various background experiences or other acoustic factors such as envelope or amplitude modulations not analyzed here. This analysis highlights certain phoneme processing methods that could apply to other forms of modified speech, paving the way for more research on whistled speech and processing methods.

2.2 Testing perceptual flexibility in speech through the categorization of whistled Spanish consonants by French speakers

Abstract

Whistled speech is a form of modified speech where, in non-tonal languages, vowels and consonants are augmented and transposed to whistled frequencies, simplifying their timbre. According to previous studies, these transformations maintain some level of vowel recognition for naive listeners. Here, in a behavioral experiment, naive listeners' capacities for the categorization of four whistled consonants (/p/, /k/, /t/ and /s/) were analyzed. Results show patterns of correct responses and confusions that provide new insights into whistled speech perception, highlighting the importance of frequency modulation cues, transposed from phoneme formants, as well as the perceptual flexibility in processing these cues.

Article Information

Article Status

This article has been published in JASA Express Letters:

Tran Ngoc, A., Meunier, F. & Meyer, J. (2022a). Testing perceptual flexibility in speech through the categorization of whistled Spanish consonants by French speakers. *JASA Express Letters*.

https://doi.org/10.1121/10.0013900

Keywords: consonant categorization, whistled speech, Silbo, whistled languages

Supplementary files:

See Annex, A.2 for SuppPub1 and SuppPub2

Testing perceptual flexibility in speech through the categorization of whistled Spanish consonants by French speakers

Introduction

Whistled speech is a naturally modified speech form characterized by the transposition of the spoken signal into whistled frequencies, drastically changing the spoken timbre. Whistled vowels of non-tonal languages are produced at relatively stable frequencies, which depend on the vowel position, the whistling technique, the language, the whistler's oral cavity and vowel coarticulation with surrounding phonemes. In such languages, the whistled F0 codes and simplifies the timbre of modal speech (See supplementary material SuppPub1 in Annex A.2, for a figure showing the Waveform and Spectrogram of a spoken and whistled Spanish sentence). Typically, /i/ corresponds to the highest whistled frequencies and /o/ to the lowest, while/e/ and /a/ are placed in the middle (with /e/ higher than /a/; Meyer, 2015). Just like in spoken speech, the consonants modify/modulate the vowel frequencies by adding stops and pitch changes as the whistlers "pronounce" the consonants while whistling (cf. Figure 1). With the added constraint of a rather closed mouth to whistle, this speech transformation generally creates an emphasis on the upper resonant cavities. This explains why whistled frequencies usually reflect frequency shapes of the second or third formants of modal speech (F2 most frequently, but also F3 for front vowels) (Shadle, 1983; Meyer, 2015). However, in the case of coarticulation between back vowels and velar/uvular consonants, whistled F0 also often resembles the F1 of modal speech (see for example /k/ in Figure 1, and see Meyer et al, 2019).

We can consider whistled speech akin to other forms of modified speech, such as speech in noise or artificial sine-wave or vocoded speech, where untrained listeners are able

to identify and categorize certain aspects, such as phonemes (Blanco et al., 2018). Whistled speech recognition and categorization experiments first started in the 1960-70s with Bearnese and Turkish whistlers, focusing on word (Busnel, 1970) and CV nonsense syllable recognition between local whistlers (Busnel et al., 1962; Moles, 1970; and see Meyer, 2015 for a reanalysis)]. In 2005, Rialland ran a behavioral experiment on whistled VCV nonsense utterances identified by a fluent Spanish whistler, obtaining 57% of correct answers with better performance for certain consonants and vowels (Rialland, 2005). More recently, Meyer et al. (2019) conducted a syllable recognition experiment (/ta/, /da/, /ka/, /ga/) with Tashlhiyt Berber whistlers. Experiments with participants who were not previously familiar with whistled speech ('naive listeners') only date back to 2005. Such studies included participants with different language backgrounds (Spanish, French, Chinese) who were tested on a whistled vowel recognition paradigm based on Spanish vowels. The results obtained were well over chance for all categories of listeners with striking differences between vowel positions and language background (Meyer, 2008; Meyer et al., 2017). These previous results make us wonder whether such a capacity, or form of perceptual flexibility allowing for phoneme categorization in spite of the reduced phonetic cues, can extend to whistled consonants. We tested French speakers' categorization capacities for whistled Spanish consonants through a behavioral experiment. The set-up of the experiment also allowed us to explore other complementary questions: does the inclusion of a training portion allow for a learning affect? Which factors underlie participants' consonant categorization?

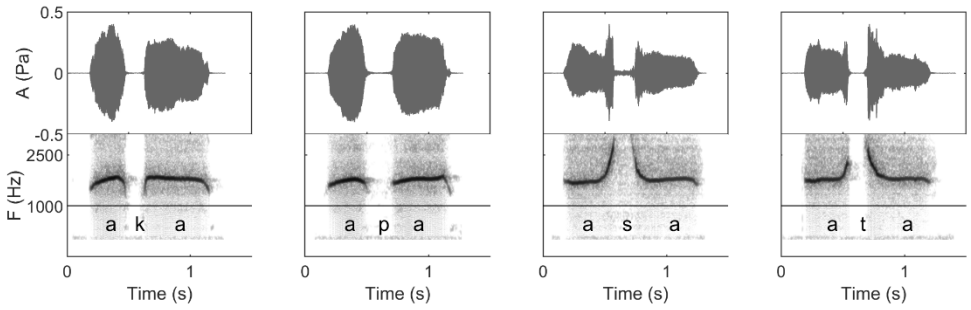
To answer these questions, we constructed our experiment in three parts including a section with whistled consonant categorization without feedback, one section with feedback and a final section with natural variations of each consonant. This allows us to test whether participants learn to apply consonant models to multiple varieties of each consonant (a

method based on a perceptual learning experiment on noise-vocoded speech, Hervais-Adelman et al., 2008). To understand the cues used for consonant categorization, we will compare the participants' responses with previous interpretations and classifications of the whistled Spanish consonant system. The whistled consonants chosen (/p/, /k/, /s/ and /t/) and recorded in Silbo (the local name for whistled Spanish in the Canary Islands), have previously been grouped in different categories based on acoustic loci, as well as frequency and/or amplitude modulations. Trujillo (2006), for example, proposed 4 consonant groups and Rialland (2005) 8 groups for all whistled Spanish consonants, both using distinctions such as "low", "acute", "continuous" and "interrupted". Both Trujillo and Rialland oppose /p/ and /k/ said to be "low" consonants, to /s/ and /t/ said to be "acute". Rialland also proposes secondary distinctions such as higher loci in /s/ than /t/ and sharper attacks in /t/. However, whistled consonant classifications are in fact more complex as they are influenced by parameters that were not systematically controlled by Trujillo and Rialland. These include whistling technique, the position of the consonant in the word, speech rate and proficiency of the whistler (Meyer, 2015). Nevertheless, all researchers agree on two clear distinctions among whistled consonants in Spanish: one between consonants with high (/s/ and /t/) or low (/p/ and /k/) whistled loci, and one between semi continuous or continuous (represented by /s/ in our experiment) and interrupted consonants (/k, p, t/). This continuous/interrupted opposition is explained by the amplitude decay either corresponding to a dip (maintaining continuity), applied to /s/, or to a complete interruption (resulting in a silence), which applies to stops. On the other hand, "acute" consonants with high loci systematically correspond to consonants rising after the previous vowel (V1) and falling towards the next vowel (V2) (see /asa/ and /ata/ in Figure 1), and low loci ("grave consonants") to the reverse (Meyer, 2015). Secondary whistled distinctions, such as those suggested by Rialland (2005), notably the sharper CV

attack, also come into play. This is typically the case of /k/ and /t/ (Meyer, 2015), whereas a more progressive CV frequency slope characterizes /p/ (see Figure 1). In this context, the classification of /s/ is one of the most complex because it emulates the continuous fricative aspect of spoken speech, expressed by an amplitude dip only by expert whistlers (see Diaz, 2017 and Figure 1). This type of whistled fricative is well termed 'semi-continuous' or 'continuous' because its acoustic continuity depends on the whistler's proficiency, the speech rate and the listening distance. In fast speech, whistled fricatives are more clearly continuous because the speed reduces the dip through a more gradual amplitude envelope modulation (Meyer, 2015); however, the low dB level present in the amplitude dip can cause its dynamics to be partly masked by the background noise in increased emitter-receiver distances.

Figure 1:

Waveform and Spectrogram of VCV forms of the experiment (whistled Spanish /aka/ (Mm.1a), /apa/ (Mm.1b), /asa/ (Mm.1c, and /ata/ (Mm.1d)[9]



Students learning Silbo from La Gomera Island generally follow recommendations based on Trujillo's groupings, however those from the Yo Silbo association, the most active Silbo revitalization association in the Canary Islands, assemble the consonants into five pronunciation-based groups using VCV configurations as a didactic basis for the classification. This classification opposes /t, s/ to /p/ and to /k/ (Diaz, 2017), where /p/ is considered low but continuous. This difference may take into account the contrast between the sharp attack

[9] See supplementary material SuppPub1 in Annex A.2 for a figure showing the Waveform and Spectrogram of the corresponding spoken utterances.

for/k/, with a more rapid release, and the softer frequency slope and amplitude modulation which characterize the consonant stop in /p/ (see Fig. 1). Even if it is not highlighted by the groupings presented above, the sharpness of the interruption could also be a defining commonality in /t/ and /k/, in opposition with /s/ and /p/. Moreover, the plosive and fricative consonant opposition is not usually proposed as a grouping characteristic, but skilled whistlers manage to develop it (Diaz, 2017). It is therefore considered as another secondarily developed opposition in the whistlers' community. To sum up, our experiment includes contrastive consonants that will shed light on the importance of these cues for categorization. Though our experiment targets a naturally modified speech form (whistled speech), the speech cues employed reflect more generalized phoneme processes as well as mental representations of phonological cues.

Methods

Stimuli

We chose to test four distinct consonants of spoken Spanish that have either identical or easily learned pronunciation differences in Spanish and French (Molina Mejia, 2007). These include three occlusive or plosive consonants ([p]-bilabial), [t]-dental/alveolar), [k]-velar) and a fricative ([s]-alveolar), followed and preceded by the vowel [a], giving the following VCV forms: [ata], [aka], [asa] and [apa]. The use of a VCV form enables us to take into account variations due to the duration of consonant closure, as well as amplitude and frequency modulations. In addition, using only one vowel reduces effects of different Consonant--Vowel coarticulations (Meyer, 2015). Four instances of each /aCa/ segment were whistled by the same proficient whistler-teacher of 'Silbo' (whistled Spanish of the Canary Islands) and

recorded by the last author. The frequencies before and after each consonant closure vary between 1141.9 and 2628.7 Hz, with an average of 1715.86 Hz[10].

Procedure and design

The experiment was programmed using PCIbex Farm and took place online from participants' own homes. Before starting the experiment, participants were asked their age, the languages they speak (and their level), as well as if they play any musical instruments. As this experiment was online, they were to indicate whether they used headphones, earbuds or speakers, and to give the name of the brand. We recruited the participants through various social media networks, excluding participants with self-declared speech/hearing impairments. Before beginning part 1 of the experiment, participants are shown the recordings of the four whistled VCV forms heard in part 1 (in a randomly chosen order) without any indication of their categorization. This allows participants to familiarize themselves (briefly) with the acoustic specificities of whistled signals as well as to adjust the volume to a comfortable listening level. The four /aCa/ recordings presented (one of each consonant, chosen according to the stability of whistled vowel frequencies, see Figure 1) are used during part 1 without any indication of the consonant heard. The participants then hear these clips in a random order and are asked to respond with either "p", "k","t" or "s" after each clip. These consonants are attributed to the arrow keys on the keyboard according to the layout of both azerty and qwerty keyboards.

Part 2 is a training phase with feedback, using the same whistled audio tracks as part 1. We first present the four different consonants in a random order by playing a spoken version of the VCV segment, followed by the whistled version. An image of the consonant

[10] See Annex, A.2, Table 2 for the each stimuli used here, presented with the acoustic signal and the spectrogram.

appears on the screen simultaneously. Following this, participants complete a shorter version of the previous test albeit with feedback. Participants hear each clip (each consonant) 4 times, amounting to 16 total excerpts. Feedback is given after each response: “Bravo” when correct and “Non ce n’était pas la bonne réponse” – “No that was not the correct answer”, when false. In part 3 of the experiment, participants hear sound clips and are requested to indicate which consonant was heard. However, in this portion, three additional versions of each consonant are included, amounting to four total variations per consonant. As this applies to all four consonants, 16 recordings are heard, out of which 12 are unfamiliar variations (i.e. not heard in part 1). Each recording is played three times and participants hear a total of 48 stimuli in part 3.

Participants

This study included 30 adults (21 women, 9 men, mean age: 29.6 years, *SD* = 8.77) whose first language was French and who did not have any language or hearing impairments. A number of participants had experience in different languages, notably in Spanish. 19 participants indicated having some experience in Spanish, where 8 participants declared being beginners, 8 participants had an intermediate level, and 3 had a confirmed level. Participants gave informed consent before starting the experiment.

Results

Our analysis focused on parts 1 and 3, excluding the short training portion (part 2) due to the small sample size. We compared both parts 1 and 3 by taking into account the 40 answers given in part 1 by each participant as well as the 48 answers given in part 3. This gave

us 3520 data components. After presenting the results, we analyzed the correct answers and the confusions separately.

When analyzing the results for correct answer percentages and confusions for the task with four possible answers, we find significantly different categorizations for the four consonants [X2(9) = 1850, *p* <.001]. Overall, the agreement of the answers with the consonant categories was different from chance and not accidental, being 'moderate' according to Cohen's kappa (k) statistics (k = 0.454, *p* <.001).

Correct answers

Participants obtained 59.2% of correct answers obtained (well above chance at 25%), i.e. participants categorized the whistled consonants properly, with the results of parts 1 and 3 pooled together. We ran a Generalized Linear Mixed Model with Spanish as a second (or third or fourth) language as a Fixed Factor and Participant as a Random effect, but found no effect. We ran a global Anova on participants with repeated measures that included Consonant type (k, p, s, t) and Part (part 1, part 3) as within factors. We observed that the scores varied significantly depending on the main effect of Consonant type (F(3,87)=16.893; *p* <.001). Meanwhile, the main effect of Part and the interaction between the two factors were not significant. This suggests that there was no significant increase in performances between Part 1 and Part 3.

Concerning consonant types, "s" and "t" obtained the largest amount of correct answers (respectively 74.5 % and 68.8%), while "k" was intermediate (52.9%) and "p" was the least well-recognized (40%). We also ran post hoc multiple comparisons with a Bonferroni correction (*p* <.05) revealing that correct "p" categorizations are significantly different from

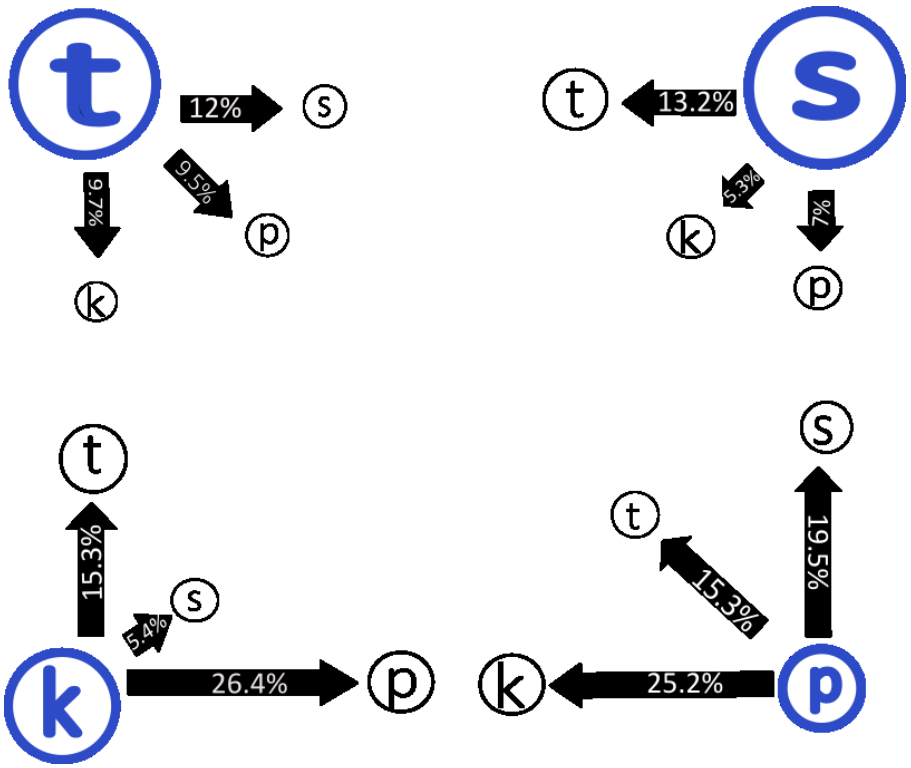
those of “t” and “s” (p <.001), that “k” correct answers are also significantly different from “s” (p <.001) and from “t” (p <.02). This opposes “p” and “k” to “s” and “t” in the following manner: “t” = “s” > “k” = “p”

Confusions

Observation of confusions in the incorrect answers allowed us to gain further understanding of the participants’ behavior. To look at confusions between consonant types we first ran a non-parametric Anova with repeated measures showing that the interaction between the two factors- Played consonant and Answered consonant - was significant (p <.005). Thus, for each played consonant, we applied a pairwise comparison (Durbin Conover) and the significant differences obtained are presented below and illustrated in Figure 2.

Figure 2:

Schematic representation of the statistical relations in the confusion matrix



The image presented shows the proportion of confusions for each consonant played (arrows), as well as the amount of correct answers obtained (by the size of the blue circle of the consonant played). The three different sizes of the smaller consonant bubbles (in black) allow us to illustrate the confusion hierarchies described thanks to pairwise comparisons.

As shown in the image presented in Figure 2, when /t/ was played, it was mistaken for “s” 12% of the cases (noted t/s), for “p” 9.5% of the cases (t/p), and for “k” 9.7% of the cases (t/k). There are significant differences between the correct answers obtained for /t/ (68.8%, noted t/t) and t/s, t/k and t/p (p <.001). These significant differences confirm that the consonant /t/ was mistaken as often for “s”, as for “p” and finally for “k”.

When the consonant /s/ was played, it was answered as “t” 13.2% of the cases (s/t), as “p” 7% of the cases (s/p), and as “k” 5.3% of the cases (s/k). Here, correct answers obtained for /s/ (74.5%, noted s/s) are significantly different from s/p, s/t and s/k (p <.001). There is also a significant difference between s/k and s/t (p <.01), indicating that /s/ was confused more often with “t” than with “k”.

When the consonant /k/ was played, it was confused with “p” 26.4% of the cases (k/p), “t” 15.3% of the cases (k/t), and “s” 5.5% of the cases (k/s). Moreover, the level of correct answers for /k/ (52.9%, noted k/k) was significantly different from k/s, k/t (p <.001) and k/p (p <.01). In addition, k/s was significantly different from k/t (p <.001) and k/p (p <.001). This indicates that /k/ was taken as much for “t” as for “p” (and more than for “s”), however the percentage of confusions presented above (Figure 2) suggest that /k/ was most often mistaken for “p”, then for “t” and finally for “s”.

Finally, when consonant /p/ (40% of correct answers, noted p/p) was played, it was confused with “k” in 25.2% of the cases (p/k), “s” in 19.5% of the cases (p/s), and “t” in 15.3%

of the cases (p/t). Moreover, p/p is different from p/t (p <.001), and from p/s (p <.01) however not from p/k. In addition, p/k is different from p/t (p <.05). This means the /p/ is most often confused with “k”, then with “s” and finally with “t”.

We can also compare the mirrored confusions (t/p and p/t; t/s and s/t; p/s and s/p; t/k and k/t and k/s and s/k), where we find a significant difference between p/s and s/p (p <.001) and a tendency between t/k and k/t (p =.09). This is not surprising due to the higher levels of consonant recognition for “s” and “t” as opposed to the low levels associated with “k” and “p”.

Discussion

The overall performance shows that participants recognized the set of consonants well over chance (25%) for every consonant type. General whistled consonant recognition averages at 59.9%, with no significant difference between the first (Part 1) and the last part (Part 3) of the experiment. In addition, there was no impact of experience with Spanish.

Though Part 3 included a greater proportion of new tokens for each consonant (75% of new stimuli), this did not affect the overall performances. This is demonstrated not only by the absence of significant effect between parts but also by the lack of interaction of this factor with the consonant type. One could have expected that categorization rate would decrease and that the new recordings would not be as well identified as the previous tokens. However, as there was in fact no difference between parts, this could suggest that participants learned consonant categorization and managed to categorize the different variations. One possible explanation is that participants learn from the consonant heard (presented in part 1 and part 2) which could act as an exemplar or instance of the phonological category. It would be

interesting to test the exact same stimuli in both parts 1 and 3 to see if a stronger learning effect can be observed.

In addition, our results show that certain consonants are easier to recognize (/s/ or /t/) and others are more difficult to recognize (/p/). The hierarchy derived from the correct answers and confusions shows a preference for the consonants with high loci, or those containing a rising pitch towards these high loci ("s" and "t", see Figure 1). This could be because the magnitude of the transitional pitch movements are greater than for the low consonants. This could also suggest that pitch movements are easier to identify than changes in articulation, envelope gaps or interruption duration.

Indeed, correct answers reflect certain preferences, reprising some aspects of previous research. In the hierarchy obtained ("s" = "t" > "p" = "k"), the preference for "s" and "t" corresponds to the opposition between "high frequency modulated whistled consonants" with high loci and whistled consonants with low loci ("k" and "p") (Busnel & Classe, 1976; Trujillo, 2006; Rialland, 2005; Meyer, 2015). The significant difference between the recognition rate of "k" and "p" from those of "s" and "t" suggests that the clear stop which characterizes our /t/, /k/ and /p/ stimuli is not a strong enough characteristic to compete with differences in saliency due to frequency slopes induced by high loci. However, the sharp attack cue, present for "k" and "t", does seem to influence perception, as /t/ is correctly categorized at 68% and is therefore differentiated from /s/.

The relative proportions of confusions reflect similarities in the perception of different consonants. Their comparison enables us to track more closely the phonetic traits to which such similarities may be due. There are three main types of traits coded in whistling: frequency modulations, amplitude envelope modulation and gap or interruption duration. For example, when looking at the main significant confusions for each consonant, we note that /s/ is

significantly more confused with “t” than with other consonants, reinforcing the interpretation derived from the correct answers obtained, and underlining that the acoustic cues associated to a high locus is key for whistled consonant categorization. Indeed, this is the principal acoustic cue that these consonants share, as /s/ is otherwise semi-continuous with a slower attack than /t/. Such a view is also supported by the high percentage of /t/ mistaken for “s” (even if, due to high variability between listeners, t/s errors are not significantly different from the errors of /t/ for “k” and “p”). Despite the fact that the confusion s/t is preferred over other confusions and the reverse is not the same for t/s, there is no significant asymmetry between s/t and t/s confusions.

The patterns of confusions of the two least well recognized consonants (/k/ and /p/) also highlight interesting phonetic aspects. As we saw earlier, /k/ is answered more as “p” than “t” or “s”, and /p/ is answered more as “k”, though there is no significant difference between the confusion as p/k and p/s. Interestingly, whistled realizations of /p/ and /t/ both share key common phonetic cues with whistled /k/: /p/, /t/ and /k/ realize a full stop, /t/ and /k/ use a sharp attack, and /p/ and /k/ share the flat frequency shape. The consonant /s/ however shares none of these characteristics with /k/. Moreover, /p/ is answered “k” and “s” at relatively similar proportions (25.12% “k”, 19.5% “s”, which are not statistically different), while “t” is answered at a significantly lower rate (15.3%, statistically different from p/k). These results may be explained by the fact that a whistled /p/ shares two phonetic traits with /k/ (full stop + flat frequency), one with /s/ (a more gradual attack than /k/ and /t/), and one with /t/ - full stop).

Overall, the results strongly confirm the hierarchy found in the correct answers: high loci (frequency shape towards high frequencies) are preferred over other phonetic cues. They also show that when several phonetic cues are shared between two consonants, this

augments their probability of confusion. However, the present study does not include enough consonantal types to classify the other key phonetic cues in whistled speech: clear silent gap, sharp/gradual attack. Interestingly, the confusion patterns also underlines the relative facility to identify /s/. Does this suggest that continuous sound with pitch change is easiest to identify in extremely modified speech?

All the results highlighted here are confirmed by the asymmetry s/p vs. p/s and the tendency t/k vs. k/t, as opposed to the symmetries k/p vs. p/k and t/s vs. s/t. Such asymmetries would be interesting to explore further with more data in the perspective of debates opened by Chang et al. (2001).

The relative ease at which /s/ is categorized by naive French listeners also contrasts with the documented difficulty for whistlers to learn to produce it. This asymmetry is all the more interesting as it may have implications for teaching whistled speech in a context of current revitalization of the practice (Diaz, 2017; Meyer, 2021). It also opens towards the possibility of convergence/divergence in production vs. perception during spoken speech acquisition (Moskowitz, 1975).

Finally, the results obtained here for this modified speech form are in line with those previously obtained by studies also dealing with whistled phoneme recognition. (a) Performance levels are coherent with those found by Meyer and colleagues for whistled vowel recognition by untrained listeners (Meyer, 2008; Meyer et al, 2017). (b) This experiment highlighted consonant preferences just as Rialland found for Silbo whistlers (Rialland, 2005). (c) Rates of correct answers + confusions were analyzed similarly to Meyer and Ridouane's analysis (Meyer et al., 2019) who also found that /t/ was better recognized than /k/for traditional whistlers of Tashlhiyt Berber (the other consonants of their test were not tested here).

Overall, with such an approach, we have shown that the naive listener capacity for recognition and categorization found in whistled vowels also applies to whistled consonants, which opens rich experimental possibilities to observe the notion of perceptual flexibility both with non-standard, but natural, whistled consonant articulations, and across different language backgrounds.

Conclusion

In conclusion, naive French listeners recognize whistled consonants above chance and generally use frequency change to identify the sound correctly, which is coherent with the fact that frequency modulations are the most salient and resilient aspects of the signal with better propagation for long distance communication. These results underline a strong perceptual flexibility present in naive listeners who can successfully identify and attribute these cues to a modified form of speech. This analysis highlights certain phoneme processing methods that could apply to other forms of modified speech, paving the way for more research on whistled speech and processing methods.

Chapter 3

Whistled vowel perception

Introduction

In this chapter (Chpt.3), we consider whistled vowel categorization by naive listeners in two contexts: the isolated vowel and within the word. In contrast with the whistled consonant (explored in the previous chapter), the whistled vowel has previously been studied in experiments testing native whistlers and naive listeners. Recent studies on naive listeners have focused on the vowels /i/, /e/, /a/, and /o/ (Meyer, 2005; Meyer, 2008; Meyer et al., 2017), and considered whistled vowel categorization by listeners with different native languages (Meyer et al., 2017), see Chpt 1.2.2.2.

This chapter, comprised of two published articles and one complementary study (in press), reprises previous studies by considering the same whistled target vowels, whilst further developing our exploration of the vowel by taking into account different forms of variability (including intra-whistler differences, inter-whistler differences, and the vowel context). These themes fall under the umbrella of a broader exploration of variability (also present in Chapter 2), which complements the behavioral studies here. In doing so, we provide more natural listening conditions which further our understanding of naive listeners' behavior when faced with whistled speech.

In the first two articles included here, we explore the effect of inter-whistler variability by including two whistlers whose whistled vowel distributions showed slight differences. Like in the consonant studies of Chapter 2, we sought to consider the effect of training on the

perception of variability in the whistled signal. To do so, a 3-part experimental design (also used in Chpt.2) was applied to these two articles. In the first article, which includes one behavioral experiment (Expt 3), the two whistlers are presented in opposing parts (one in part 1, and the other in part 3). In the second article, the complementary study, we also introduce a behavioral experiment (Expt 4) which completes the previous article (Expt 3) by including participants who hear only one whistler throughout the three-part experiment. Thus, taken together, these two experiments consider the effect of whistler variability and the effect of training on the integration of such variability. In the first article, we also contrast two different forms of participation: online and in person, by including participants who completed the experiment in both of these conditions.

In the third article included in this chapter, we focus on the whistled vowel in the context of the disyllabic word, first taking time to describe the differences produced in whistled vowel production due to coarticulation in the word. The experiment included here contains only intra-variability in the whistler productions, as it does not compare whistlers. These descriptions complement the proposed exploration of the whistled vowel in isolation (Expts 3 & 4), where the vowel stimuli were extracted from CV pairs with various consonants (/d/, /g/, /k/, /t/). Here, however, we consider the role of the vowel's position in the word, rather than the consonantal coarticulation. The behavioral experiment included in this third article (Expt. 5), maintains the focus on target vowels. It does so by analyzing whistled words with target vowels equally distributed in both positions. This also allowed us to consider the effect of vowel to vowel co-articulation in the word, by analyzing the interval created between vowels and its effect on perception. Because this third article focuses on the whistled word, the experimental design differs from that of the two previous articles in this chapter and does

not include a training portion. As such, this experiment contains a single part. These articles and their experiments are presented in Table 5.

Table 5:

Description of the experiments in the articles of Chapter 3

Chapter 3 - Vowels	Article	Expt	Target	Design	Participants	Location
3.1	Tran Ngoc et al., 2020b	Expt 3	Whistled vowels (2 whist)	3 parts	37	Online + In person
3.2	Tran Ngoc et al., 2023e	Expt 4	Whistled vowels (2 whist)	3 parts	44	Online
3.3	Tran Ngoc et al., 2023a	Expt 5	Whistled words	1 part	19	Online

3.1 Whistled vowel identification by French listeners

Abstract

In this paper, we analyzed whistled vowel categorization by native French listeners. Whistled speech, a natural, yet modified register of speech, is used here as a tool to investigate perceptual processes in languages. We focused on four whistled vowels: /i, e, a, o/. After a detailed description of the vowels, we built and ran a behavioral experiment in which we asked native French speakers to categorize whistled vowel stimuli in which we introduced intra- and inter- production variations. In addition, half of the participants performed the experiment in person (at the laboratory) while the other half participated online, allowing us to evaluate the impact of the testing set up. Our results confirm that the categorization rate of whistled vowels is above chance. They reveal significant differences in performance for different vowels and suggest an influence of certain acoustic parameters from the whistlers' vowel range on categorization. Moreover, no effect or interaction was found for testing location and circumstances in our data set. This study confirms that whistled stimuli are a useful tool for studying how listeners process modified speech and which parameters impact sound categorization.

Article Information

Article Status

This article has been published in the proceedings of INTERSPEECH 2020:

Tran Ngoc, A., Meyer, J. & Meunier, F. (2020b). Whistled vowel identification by French listeners, *INTERSPEECH 2020 – 21th Annual Conference of the International Speech Communication Association, September 14-18, Shanghai, China, Proceedings*, 1605-1609. https://doi.org/10.21437/Interspeech.2020-2697

Keywords: vowel categorization, whistled speech, whistled languages, speech perception, acoustic cues

Acknowledgements

We wish to thank the two whistlers for their contribution in whistling words, Jonathan Parente for his help with the vowel experiments in the lab and all the participants for volunteering their time.

Whistled Vowel Identification by French listeners

Introduction

Whistled speech is a type of natural speech, which transposes spoken speech into whistles (see Meyer, 2015 for a review). At least 40 low-density and remote populations have adapted their local language to this particular speech modality, using it for long distance communication. Notably, whistled speech is intelligible only to trained speakers, and is not directly comprehensible to naive listeners even if they are fluent in the language that is being whistled (Busnel & Classe, 1976).

Transposition from spoken speech to whistled speech in most non-tonal languages relies on a 'formant-based whistling strategy' (Meyer, 2015). Whistlers make an approximation of the vocal tract articulation used in the spoken form to pronounce the whistled phonemes. In Spanish for example, whistled vowels are emitted at different pitch levels depending on the frequency distribution of the whistler's timbre in the spoken modal speech form (i.e., /i/ has a high pitch, /e/ lower, /a/ even lower, and /o/ the lowest; Meyer, 2008).

Previous studies on whistled speech have proved that naive listeners recognize whistled phonemes using acoustic cues. A first experiment conducted in 2008 showed, using different productions from a single whistler, how naive French listeners were able to categorize whistled Spanish vowels /i, e, a, o/ with a mean level of success corresponding to 55% of correct answers (Meyer, 2008). In 2017, a second experiment using the same stimuli showed that the scores varied per vowel: /a/ and /e/ showed the lowest scores (44.1 and 46.9%, respectively), and /o/ and /i/ were recognized best (50.6 and 78.4% of correct categorizations, respectively, with /i/ being significantly different from the other vowels). The

authors also took an interest in the impact of listener experience on vowel recognition, finding that one's native language (Spanish, French or Standard Chinese) impacted whistled vowel categorization, though the results of the French and Spanish participants were not significantly different (Meyer, 2008; Meyer et al., 2017).

While previous studies on whistled speech have included some intra-talker variability, very few studies have addressed this variability in whistled speech, despite research showing that inter-talker variability in noise has significant effects on spoken speech perception (Zaar & Dau, 2015). In addition, a correlation between certain acoustic phonetic properties and listener comprehension has been observed for non-native listeners (Bent et al., 2010): talkers with a larger vowel space were, indeed, easier to understand. An experiment displaying a combination of these conditions (native and non-native listeners with inter-talker variability and presented in slight noise) showed similar results with a significant effect of inter-talker variability on intelligibility (Dommelen & Hazan, 2012). These properties, different for native and non-native listeners, include more energy in the 1-3 kHz range, as well as an enlarged vowel space in the F2 range. Interestingly, the stimuli from these experiments deal with certain constraints which also characterize whistled speech (modified speech forms that are first unintelligible for naive listeners) leading us to investigate the impact of acoustic phonetic inter-talker variations in whistled speech perception.

The present paper extends the previous experiments on whistled speech while considering the impact of slight inter- talker variation with several objectives. First, it aims at (a) testing whistled vowel categorization with new whistled stimuli, to assess whether the previous results can be generalized. It then (b) seeks to introduce inter-individual differences (inter-talker variability) in the productions tested, using stimuli from two different whistlers.

It also (c) explores the possibility of a learning effect throughout the different parts of the experiment using a transfer-learning model (Wang & Zheng, 2015) and finally (d) looks at the impact of the testing set up by comparing data acquired in the lab with data obtained online with participants running the experiment from home. This is particularly relevant to the current quarantine period, which prevents many researchers from conducting experiments in laboratories.

To answer these questions, we constructed a three-part experiment. Part 1 asks participants to respond to stimuli without any previous introduction, part 2 proposes a short learning phase where feedback is given, and finally part 3 consists of the same test as part 1, with stimuli from the other whistler. This allows us to evaluate learning by comparing parts 1 and 3. Finally, to test for potential effects of the experiment set up, half of the subjects participated in the experiment in the lab and the other half participated online from their homes.

Experiment

Method

Stimuli

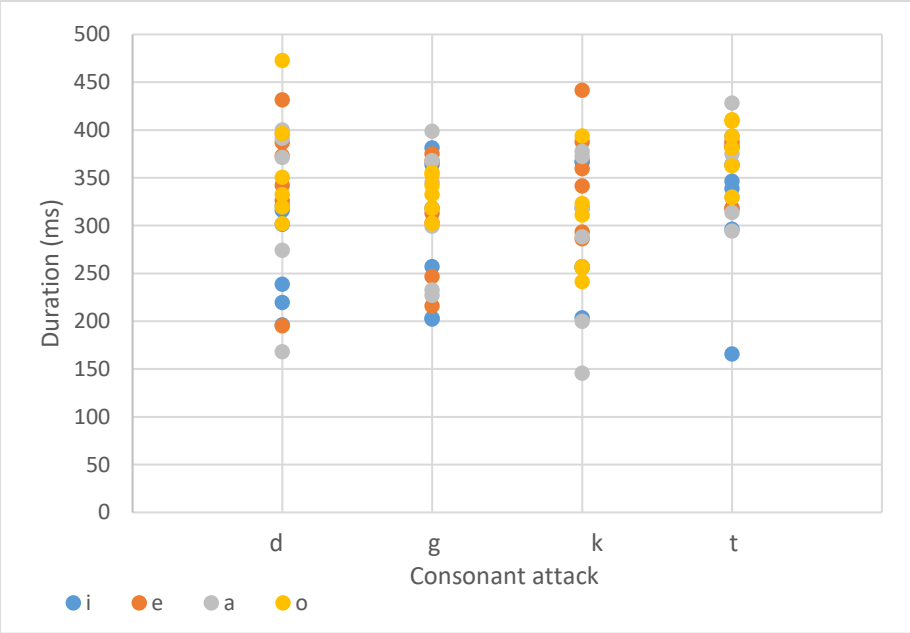
This experiment was conducted in accordance with the Helsinki agreement. Julien Meyer recorded the stimuli in a soundproof room of the Gipsa-Lab (Bedei Platform) with two different expert whistlers, both teachers of whistled speech in the Canary Islands. The whistled Spanish vowels /i/, /e/, /a/ and /o/ were extracted from bisyllabic CVCV whistled words (such as /cada/, /nata/…). In order to retain the same prosody for each vowel chosen as a stimulus for the test, we systematically selected vowels from the second CV syllable, on

unaccented syllables only. Moreover, we selected vowels following various consonant attacks (/d/, /k/, /g/, /t/), and, after removing the consonant attack, silence was added to the vowels to create homogenous samples of 500 milliseconds.

The extraction of these whistled vowels from CVCV words causes their duration to vary a great deal. As that duration can be discriminated easily for any difference over 100 milliseconds (Gelfand, 2016), we chose to use whistled excerpts of sufficiently varying lengths (see Figure 1) to ensure that the overall duration differences between the stimuli could not be used to discern the individual vowels. The vowel stimuli therefore last between 146 ms for a whistled /a/ extracted from a /ta/, to 473 ms for a whistled /o/ extracted from a /go/ (both by whistler A). These durations vary according to the vowel, the whistler producing the stimuli (the recordings of whistler A vary more in duration than those of whistler B) and the consonant attack.

Figure 1:

Whistled vowel duration following consonant attacks



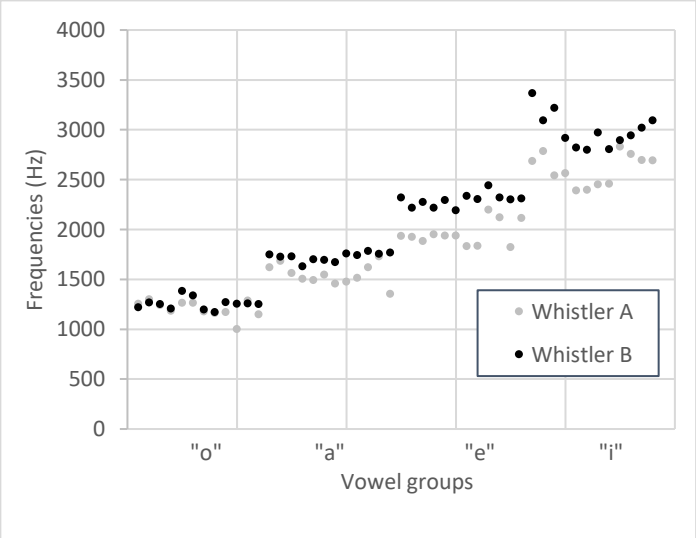
In addition, the durations loosely reflect those used in an experiment on vowel length in English (Everest & Pohlmann, 2009). In this experiment, synthesized versions of natural

vowels were created with 3 different durations: 272, 144 and 400 ms. Results showed that duration had a small overall effect on vowel identification.

The frequency of the vowel also varies (see Figure 2). This variation is slightly influenced by the consonant attack (as seen in Figure 1) which was removed from the recording and then replaced by a fade-in. However, variation is generally attributed to factors such as the whistlers' physical morphology, vocal range, whistling technique employed for producing the whistle (see Meyer, 2015, for a review), and whistling skills. In line with the previously mentioned experiments (Bent et al., 2010; Dommelen & Hazan, 2012), whistler B, who often teaches Silbo, believes that the further apart the vowel groups are, the easier it is to distinguish and identify them. This also echoes the tendency languages have for maximizing acoustic distances between vowels, often described in linguistic theory (Hillenbrand et al., 2000). Here, we observed that the frequencies of different vowel positions of whistler A are proportionately less spread out than those of whistler B (Figure 2). This applies to all of the vowels except for [o], where the difference is approximately 50 Hz which can be attributed to morphological variations between the whistlers.

Figure 2:

Distribution of whistled frequencies according to vowel groups per whistler. Whistler A productions are the darker dots.



Certain vowel groups vary more than others: this is especially the case for high frequency whistled vowels, /i/ and /e/ (see Figure 2 and Table 1). In contrast, /a/ and /o/ are more stable for both whistlers (Figure 2 and Table 1). Overall, not only are the vowel groups of whistler B more distanced from one another, the frequencies are also more stable, reflecting the use of different whistling strategies.

Due to the variation of duration and frequency as well as the importance of relative frequency perception in whistled speech (which relies on modulations of a simple frequency line) participants may need to identify the "range" of the vocalic whistled space of the whistler, which remains proportionately uniform for each individual. The relationship between the vowel frequencies presented in Figure 2 allow us to deduce two linear equations (derived from the values obtained from the linear regression on all the data attributed to each whistler). These equations, based off the average frequency of the vowel group for each whistler, underline the difference in slope and of whistled range, which become more important for the vowels /e/ and /i/. In the linear equations below we considered *x* to be the position of the vowels, following the order [o, a, e, i] where [o]=1, [a]=2, [e]=3 and [i]=4, and *y* to be the average vowel frequency. This distribution and vowel order also reflects those of the French or Spanish vowel diagram (or triangle) starting from "back" and "closed" and moving towards "front" and "open". Equation 1 corresponds to whistler A and equation 2 to whistler B.

$$y = 460.9x + 677.01 \quad (1)$$

$$y = 578.37x + 622.36 \quad (2)$$

Despite the difference in slope between the two equations, the general relationship between the vowel groups is around 4/3 of the average frequency of the vowel below it, though slightly lower for whistler A and slightly higher for whistler B.

In this experiment, we maintained the relationship between the whistler and the whistled vowel range by testing the whistlers separately, taking into consideration a possible effect of inter-talker variability and testing whether participants adapt to an individual whistler-specific frequency distribution or a general frequency distribution. In addition, this reflects a more realistic situation concerning the ecological conditions: when whistlers hear each other, they adjust to the other person's range to understand their speech.

Design

We evaluated how naive participants performed on categorizing whistled vowels using one whistler's productions and, after a training section using the vowels of the same whistler, we evaluated how these performances changed when responding to the other whistler's productions. This procedure enabled us to test whether there was an overall learning effect from listening to the first whistler (transfer-learning model), or whether listeners rely on other parameters such as relative frequency perception (similar to the perception of musical notes). This experiment has two versions (one with whistler A first and one with whistler B first) both containing three parts, i.e. part 1 (test), part 2 (training) and part 3 (test).

In part 1, participants listen to 48 whistled vowels, corresponding to 12 versions of each vowel type. These include 3 different recordings of each vowel extracted from the same consonantal context. Part 2, the training session with feedback, comprised 16 vowels, using 4 recordings of each vowel, each corresponding to a different consonant attack. We chose these

recordings from the 48 heard in part 1 according to their proximity with the average frequency of that vowel (Table 1).

Table 1:

Average frequency (m) and standard deviation (SD) of vowels according to whistler

	Average frequency of vowels (Hz)			
Whistler	"i"	"e"	"a"	"o"
m (A)	2605.02	1958.59	1547.82	1205.61
SD (A)	156.27	123.19	104.59	82.94
m (B)	2995.16	2294.59	1726.85	1256.51
SD (B)	173.98	66.16	44.61	58.79

In part 3, participants listen to the stimuli from the other whistler which consist of 48 whistled vowels (12 versions of each vowel type, with the same criteria as part 1). If participants created an abstract representation of the vowel during parts 1 and 2, they should be able to recognize the stimuli from part 3 better than those from part 1.

The online experiment was programmed with PCIbex Farm using headphones, earbuds or speakers at home. The in-person experiment took place in a quiet room in the BCL lab (MSHS, Nice, France), was programmed using PsychoPy, and used Sennheiser HD 200 Pro or Sennheiser MB360 headphones. All other parameters were identical.

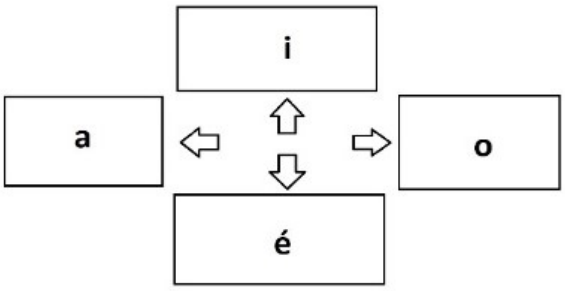
Procedure

Before starting the experiment, we asked participants to indicate the languages they speak and their musical experience. In the online version, participants informed us whether headphones, earbuds or speakers were used, and the corresponding brand. Online, participants were to adjust the volume to a comfortable listening level, in person however, we set the headphones at a fixed comfortable volume.

Part 1: This part presents participants with recordings performed by one of the whistlers. It asks participants to categorize the whistled vowels heard without any training using the arrow keys. The arrow keys are attributed to each vowel following the keyboard layout (both qwerty and azerty), and are presented before and during the experiment (Figure 3).

Figure 3:

Arrow keys assigned to each vowel



Part 2: Participants then complete a short training session with feedback for 4 versions of each whistled vowel. If the participants heard whistler A in part 1, the training used whistler A's recordings. If they heard whistler B in part 1, the training used whistler B's recordings.

Part 3: Finally, participants are asked to categorize the whistled vowels of the other whistler (if they heard whistler A in parts 1 and 2 now they will hear whistler B and the reverse if they first heard whistler B). Aside from using the other whistler's recordings, this part is identical to part 1.

Participants

Thirty-seven participants were tested for this experiment; they were all native French speakers aged between 19 and 50 years old (M = 26.8; SD = 8.37). They did not have any language or hearing impairments and did not play any instrument at a high or pre-professional level. Participants gave informed consent before starting the experiment. Seventeen

participants completed the experiment in the lab and the other 20 participated online. We recruited the participants online through various social media networks, and in person through the University Côte d'Azur, considering that, once we excluded self-declared speech/hearing impairments, participants did not have any pre-disposed differences in performance.

Results

In our analyses we took into account the 48 answers given in part 1 and the 48 answers given in part 3 by each participant. Overall, we obtained 53.5 % of correct categorizations out of the 3352 answers given. We first ran a global repeated measures Anova that included 2 within fixed variables -Vowel type (/a,e,i,o/) and Part (part 1, part 3)- and 2 between subjects fixed variables: Order of presentation (whistler A first, whistler B first) and Experimentation (online, in the lab). We considered the Participant factor to be random.

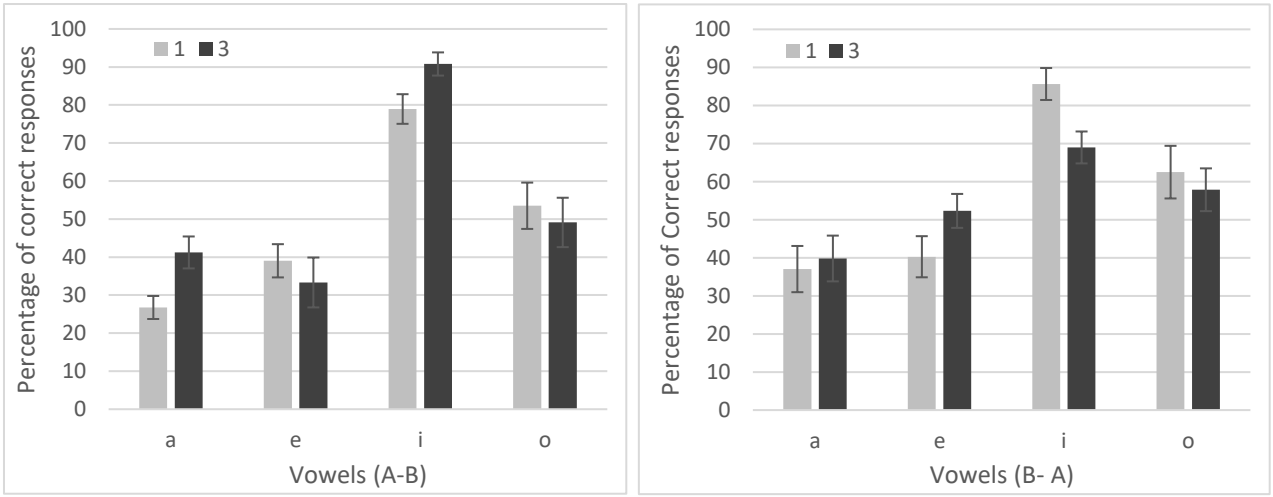
We first noted that Experimentation (online or in the lab) is never significant (at threshold of .05): neither alone, nor in interaction. We observed a significant effect of Vowel type ($F(3,96)=59.594$; p <.001). It appears that /i/ is categorized correctly 86.04% of the time, /o/ 58.95%, /e/ 43.56% and /a/ 38.31%. The interaction Vowel type * Order of presentation * Phase is also significant ($F(3,96)=3.520$; p =.02). In order to understand this double interaction we ran two other Anovas, one for each Order of presentation.

When the productions of whistler A are presented in part 1 and the ones of whistler B in part 3 (see results in left-hand graph of Figure 4), we observed an effect of Vowel type ($F(3,51)=44.53$; p <.001) as well as a significant interaction between the Vowel type and Part ($F(3,51)=3.82$; p=.015). We then ran a post hoc test to look at specific comparisons and used a Bonferroni correction (p <.05) in order to perform a multiple comparison test. It appears

that specific comparisons between each vowel with itself in parts 1 and 3 are not significant, showing that there is no specific learning for one particular vowel. Within part 1 we observed that /i/ is better categorized than the 3 other vowels /o, e, a/ and that /o/ is different from /a/ which was hardest to categorize correctly. In part 3, significant differences (p <.05) were observed only between /i/ and the 3 other vowels /o, e, a/.

Figure 4:

Correct whistled vowel categorization when the productions of whistler A are presented in part 1 and the ones of whistler B in part 3 (chart on the left) and when (chart on the right) the productions of whistler B are presented in part 1 and those of whistler A in part 3.



When the productions of whistler B are used in part 1 and the ones of whistler A in part 3 (see results in Figure 4, right-hand graph), we observed an effect of Vowel type ($F(3,48)=21.391$; p <.001) but no significant interaction between the Vowel type and Part. Running a post hoc test with the Bonferroni correction to look at specific comparisons (p <.05), it appears that /i/ is better categorized than the 3 other vowels /a, e, o/ and that /a/ is different from /o/. The difference between /e/ and /o/ also shows a tendency (p =.07).

Discussion

In this experiment, we looked at how whistled vowels are categorized by naive listeners. We aimed at extending previous results to see if they can be generalized, and introduced inter-individual differences (talker variability) in the productions to see how abstract the representations stored in the brain are, and if certain acoustic phonetic cues allow for better phoneme perception. In addition, we checked for a learning effect throughout the different parts of the experiment. Finally, we explored the effects of online or in-person testing set up.

Overall, whistled vowel categorization was obtained with 53% of correct responses (well over chance 25%), confirming the results obtained previously (Meyer, 2008; Meyer et al., 2017). Having used stimuli from two different whistlers in this experiment, the previous results can be generalized, as they also apply when participants are faced with natural variations of whistled vowels. The vowel specific differences were also replicated (Meyer, 2008; Meyer et al., 2017), where /i/ was categorized best and was systematically different from the others vowels, followed by /o/, /e/ and /a/ for which /e/ and /a/ were harder to recognize and were not different from each other. This generalization is also supported by the lack of significant difference found between the results of online and in-lab participants: whistled phonemes are recognized equally well in the two conditions.

The inter-talker variation between whistlers also proved to have an impact, as suggested by the interaction observed in the global analysis. When whistler A (with a smaller vowel space) was presented in part 1, /i/ was better recognized than the other vowels /o, e, a/, and /o/ was distinctive from /a/. Yet this was not the case in part 3 for whistler B where only /i/ was better recognized than the other vowels. When whistler B (with a larger vowel space) was presented first, there was no difference between parts, /i/ being recognized best

and /o/ being distinctive from “a” both in parts 1 and 3. In addition, /e/ and /o/ showed a tendency to be different. This suggests that when participants heard whistler B first, the abstract representation of sounds was more easily applicable to whistler A not only for /i/ distinctions, but also for /a/ (and /e/) distinct from /o/. When whistler A was first, these representations only applied to /i/ different from /e/, /a/ and /o/. In line with existing literature (Dommelen & Hazan, 2012), our findings suggest that more stable frequencies and larger vowel space facilitate abstract representations of the middle vowels (/e/ and /a/).

Finally, there was no overall learning effect found, though there were some significant differences for specific vowels. This shows that the training portion (part 2) did not systematically help construct an abstract representation of the sounds heard. To better test for the creation of abstract representation, further experiments should be conducted with the same whistler in parts 1 and 3. In addition, to better measure the effect of talker variability, more whistlers should be included in future experiments.

Conclusions

In conclusion, naive French listeners recognize whistled vowels between 53 and 55% of the time. These results appear to be robust and generalizable. Our study further showed that the whistler’s range and frequency distribution influenced participants’ categorization of vowels, and that larger vowel space facilitates the creation of abstract vowel representations.

3.2 Whistled phoneme categorization: the vowel space range effect

Abstract

We explore whistled vowel categorization by untrained listeners, focusing specifically on the impact of the different vocalic frequency ranges of two whistlers (for the vowels /i/, /e/, /a/, /o/) and the effect of training on performance. In the experiment, we included stimuli that show inter-individual and intra-individual variations of production. In the analyses we looked at the whistler identity effect and at the learning effect through the experiment for the studied vowels. The results showed an effect of the whistler, where the larger vocalic range led to improved categorization, and highlighted the robustness of the vowel recognition hierarchy. There was no general learning effect, albeit for one vowel and for the whistler with a narrower vocalic range. This study provides insight into one's representation of the vowel space in non-tonal languages.

Article Information

Article Status

In Press – ExLing Conference Proceedings

Tran Ngoc, A., Meyer, J. & Meunier, F. (2023e). Whistled Phoneme Categorization: the Vowel Space Range Effect. *ExLing Conference Proceedings Athens*.

Whistled Vowel Categorization: the Effect of Vowel Space and Training with Feedback

Introduction

Whistled speech is a natural speech form used for long distance communication. To do so, it transposes spoken speech into whistles produced in the front oral cavity of the mouth. In non-tonal languages, vowels are emitted at different whistled pitch levels depending on spoken vowel qualities (Busnel and Classe, 1976). For example, in Spanish, whistled /i/ has the highest mean values of pitch, /e/ is lower, /a/ is even lower, and /o/ even more so (Meyer, 2008). While whistled speech is not directly understood by naive listeners - i.e. listeners who never heard it before - previous studies have proved that they categorize whistled vowels much better than chance (Meyer et al, 2017, Tran Ngoc et al, 2020). In the present experiment we used whistled speech as a tool to investigate perceptual processes in language processing, more specifically to test the impact of production variations in the Vowel Space Range.

Experiment

Methods

We ran a behavioral experiment in which we asked 44 naive participants (French-language natives) to categorize whistled vowel stimuli. We focused on four whistled vowels: /i, e, a, o/ whistled by two different whistled Spanish teachers in the Canary Islands: whistler A had a more restricted vocalic frequency range, and whistler B had a wider range (see Tran Ngoc et al., 2020b for details). Stimuli were extracted from the stable whistled vowel nuclei of the second vowel of CVCV words (such as /cada/, /nata/…) following various consonants to

introduce variations (/d/, /k/, /g/, /t/). The experiment was structured in 3 parts. In part 1, participants listened to 48 whistled vowels (12 versions of each vowel type). Part 2 was a short training session with feedback (16 stimuli) produced by the same whistler as in part 1. Part 3 was similar to part 1. Stimuli were presented in a random order in each part. Four versions of the experiment were built, called AA, BB, AB, BA, according to whether productions of whistler(s) A and/or B were presented in parts 1 and 3.

Results

General Aspects

We took into account the answers given in part 1 and in part 3 by each participant. We find that overall, the 44 participants obtained 53.55% (SD = 12.99) of correct responses out of the 2112 answers given (well over chance, at 25%). We compared different conditions of the experiment by running various Generalized Linear Mixed Model analyses, described below. When convenient, the post hoc tests (all with Bonferroni corrections) are summarized by the symbols > or =, respectively indicating a significant difference or no difference.

Comparison between AA, BB, AB and AB conditions

In a first analysis, we looked at the effect of having either only one whistler or two in each list (throughout parts 1 and 3). Taking into account the whole set of data (the 44 participants), we ran a GLMM on Correct Answer with Part (1,3), Whistler identity (A, B) and the Number of whistlers per list (1 or 2) as fixed factors and Participants as a random factor. We find a significant main effect of Whistler identity (X^2 (3, N=44) = 5.9505, p= .01) as well as a significant interaction between Whistler identity and Number of whistlers in the lists (X^2 (3, N=44) = 6.8105, p< .01). The post hoc tests revealed a difference between whistler A and

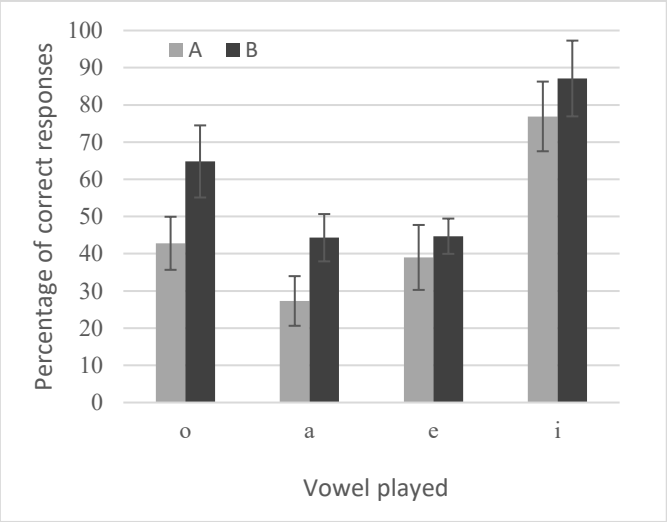
whistler B only in the comparison between the lists AA and BB where the same whistler was included in both parts (p< .05).

Correct answers for the AA/BB experiment

Considering only the 22 participants who heard the versions with only one whistler (either AA or BB), we applied a GLMM on Correct Answer, with Part (1, 3), Vowel (/i, e, a, o/) and Whistler identity (A, B) as fixed factors and Participants as a random factor. We observed significant differences between the vowels, (X2(3, N=22)=247.48, p< .001) (where /i > o > a=e/) and a significant effect of Whistler identity (X2(3, N=22)=6.10, p= .014) showing that whistler B's productions give rise to much better performances than whistler A (60.23% vs. 46.49%). A significant interaction Vowel*Part (X2(3, N=22)=21.62, p< .001) revealed that no vowel showed a significantly better performance in one part compared to the other, though there were differences in vowel recognition hierarchies between parts: /i > o (=a) > e (=a) /in part 1; and/i > o = e > a/in part 3. Finally, the significant interaction Whistler*Vowel (X2(3, N=22)=7.99, p< .05) showed that for whistler A the hierarchy was /i > o (=e) > a (=e)/; whereas for whistler B /i>o>e=a/ (see % of correct responses in Figure 1). These results suggest a stable hierarchy between /i > o > a/, with /e/ being less stable and suggesting that, through the experiment with the productions of whistler A, /e/ is better categorized.

Figure 1:

Correct whistled vowel categorization per whistler (in %)



Correct answers analyzed according to Whistler (A,B) for all lists

To gain statistical power and considering we didn't observe an effect of part according to whistler in the previous analysis, we took into account the whole data set (44 participants) and applied the same type of GLMM analysis, thus looking at the global picture. We found a significant main effect of Vowel (X2(3, N=44)=515.02, *p*< .001) as well as three significant interactions between Whistler and Vowel (X2(3, N=44)=32.36, *p*< .001), between Part and Vowel (X2 (3, N=44)=7.93, *p*< .05), and a double interaction Whistler*Vowel*Part (X2 (3, N=44)=11.04, *p*< .05). Post-hoc analyses showed that for whistler A (see Figure 2), the only vowel showing a better performance in part 3 than in part 1 is /e/ (*p*< .01). The hierarchy found in part 1 is /i > o > a = e/; whereas in part 3 it is /i > o = e > a/. Post-hoc analyses also revealed that for whistler B no vowel showed a significant difference in performance between parts 1 and 3 and the hierarchy is similar in both parts: /i/ > /o/ > /e,a/. Moreover, the only specific comparisons that reach significance comparing the two whistlers are for /e/ and /i/, both in

part 3, (respectively $p < .01$ and $p < .05$), in which /i/ is better recognized by listeners while hearing whistler B and /e/ is better recognized while hearing whistler A.

Figure 4:

Correct whistled vowel categorization (per vowel, part and whistler)

Discussion and Conclusion

This experiment first shows that the range of the vowel space used by different whistlers affects vowel categorization. The whistler with the wider vocalic frequency range gave rise to the best categorization rates, in line with the literature showing that hyper-articulation improves speech processing and that expanded vowel space benefits listeners, both natives and L2 speakers, in silence or in noise (Kagantharan et al. 2022). Moreover, learning through the experiment appeared restricted to only one vowel for the whistler with the narrower frequency range. Interestingly, this vowel /e/ has the least spoken formant convergence, which could explain less stability in recognition (Chistovitch & Lublinskaya 1979). Overall, the results highlight the robustness of the vowel recognition hierarchy previously observed, and a certain stability in the speech perception process when faced with inter-talker variability.

3.3 The Effect of whistled vowels on whistled word categorization for naive listeners

Abstract

In this paper, we explore whistled word perception by naive French speakers. In whistled words of non-tonal languages, vowels are transposed to relatively stable pitches, which contrast with consonant movements or interruptions. Previous studies on whistled speech with naive listeners have tested vowels and consonants separately. Other studies on spoken word recognition have found that vowels and consonants contribute differently to intelligibility, where the role of vowels was highly mediated by the context. Here, naive participants recognize disyllabic whistled words above chance, and vowels are shown to contribute differently than consonants. When focusing on the role of vowels, we found different scales of performance between the vowels tested, mediated by their position in the word. We also highlighted the importance of the vowels' relative frequency difference (called 'interval') in the word.

Article Information

Article Status

This article has been published in the proceedings of INTERSPEECH 2023:

Tran Ngoc, A., Meunier, F., Meyer, J. (2023a). The Effect of Whistled Vowels on Whistled Word Categorization for Naive Listeners. *Proc. INTERSPEECH 2023*, 3063-3067.

https://doi.org/10.21437/Interspeech.2023-1967

Keywords: speech perception, whistled speech, word perception, vowels

The Effect of Whistled Vowels on Whistled Word Categorization for Naive Listeners

Introduction

Speech perception is a complex process that requires great flexibility, especially when the speech form is modified or difficult to hear. Here, we take an interest in whistled speech, a naturally modified speech modality which transposes spoken words to a simple melodic line within the highest functional frequencies of the voice spectrum (~ 1 - 4 kHz).

This form of speech, used in mountainous and forested regions to communicate over long distances, transforms the speech signal into a whistled pitch modulation according to certain aspects of modal (spoken) speech (Busnel & Classe, 1976; Meyer, 2021). In most non-tonal languages that use whistled speech, the vowels are transposed to relatively stable whistled frequencies, which also depend on factors such as speaker, whistling technique, and coarticulation with the surrounding phonemes. In whistled Spanish, used in the Canary Islands, the mean whistled pitches of the 5 Spanish vowels were found to be ordered from highest to lowest as /i/, /e/, /a/ and /o/, with /u/ generally overlapping with /o/ and sometimes with /a/ (Rialland, 2005; Classe, 1956; Diaz, 2008; Meyer, 2015). Whistled consonants modulate these pitches through their corresponding spoken articulation. This can cause rapid pitch changes (for example for consonants /s/ and /t/) or stops at stable pitches (for example for /k/ and /p/) (Diaz, 2008; Meyer, 2015).

Here, we seek to study French whistled word recognition by naive listeners (listeners who are unfamiliar with whistled speech). As whistled speech conserves essential components present in modal speech, trained whistlers manage to reach high levels of intelligibility without

being restricted to certain words or set phrases. Though this form of speech is fully comprehensible to native whistlers (with natural conditions and repetition, sentence comprehension may reach 100%), in psycholinguistics tests, sentences are usually understood by trained listeners between 70-80% of the time (Meyer, 2015; Moles, 1970). Word perception in previously performed tests by Busnel showed an identification rate at around 60-75% (for 40-50 words in whistled Turkish) [8]. These whistled word identification rates show a 20-30% increase in correct responses when compared to tests based on CV or VCV tokens (Meyer, 2015).

So far, whistled word identification with naive listeners has not been studied. However, in 2020 and 2022, several experiments conducted on whistled speech perception showed that naive French listeners' perception of whistled phonemes is well over chance (Tran Ngoc et al., 2020a; Tran Ngoc et al., 2020b; Tran Ngoc et al., 2022a). In these four alternative forced-choice studies (4-AFC), listeners showed similar categorization rates between vowels and consonants, though there were preferences for certain consonants and vowels among the four that were tested. The categorization rates were organized as follows: /i/ > /o/ > /a/ = /e/ and /s/ = /t/ > /k/ = /p/.

Studies on modal speech degraded by noisy conditions show that vowels are, in general, far better recognized than consonants when they are presented in words or non-words (Meyer et al., 2013; Benki, 2003; Varnet et al., 2012). Indeed, they are more salient in adverse conditions because of their energy and stability. For this reason, they play an important role in the step preceding lexical identification: detecting the word (Meyer et al., 2013). In general, the contribution of vowels to word recognition is mediated by context. For example, Fogerty & Humes (2010) show that the vowels of monosyllabic words facilitate

intelligibility in sentences much more than in isolated words, whereas consonants are used equally in both contexts. The relationship between vowels is also found to be a contributing factor for word recognition in specific cases, such as in CVCV words (Delle Luche et al., 2014).

Therefore, we wonder how naive French listeners will use vowels and consonants to recognize disyllabic words whistled in their own language. Moreover, we wonder how whistled vowel categorization will affect whistled word recognition, not only because of the differences in categorization rates found in isolated whistled vowels (i), but also because of the important role of adjacent vowels found in modal speech (ii) (Delle Luche et al., 2014). Concerning (i), the whistled vowel recognition rates obtained in previous studies (Tran Ngoc et al., 2020a; Tran Ngoc et al., 2020b; Tran Ngoc et al., 2022a) shows that the vowels at both extremities of the whistled pitch range were categorized best. As for (ii), perceptual tests on vowels have already found that confusions between vowels presented one after the other occurred mostly between frequency neighbors, and that a significant frequency jump (i.e. a larger relative inter-vowel frequency interval) reduced the confusion rates. Thus, we hypothesize that participants may have more facility with words containing larger inter-vowel intervals, particularly when including the highest and lowest whistled vowels (/i/ and /o/).

As whistled word recognition has never been tested previously with naive listeners, we sought to maintain continuity with previous whistled phoneme experiments. We chose whistled words enabling us to target the whistled vowels and consonants used in previous experiments (Tran Ngoc et al., 2020a; Tran Ngoc et al., 2020b; Tran Ngoc et al., 2022a). By presenting these words in a disyllabic C1V1C2V2(C3) form, we can test these vowels in different contexts according to their position in the word and inter-vowel interval. To sum up, the aims of this study are first to test naive listeners' capacity for whistled word recognition.

Next, we take an interest in the role of whistled vowels in comparison with consonants in this task. Finally, we explore the differences between vowels in different positions (V1 and V2) as well as the effect of the relative vowel frequency interval on whistled word recognition.

Experiment

Method

Stimuli

We included 24 French words in this recognition task. These words were selected to integrate the target vowels and consonants from previous experiments.

The selection criteria includes the following:

- The selection of disyllabic nouns with the following structure: CVCV(C), noted as C1 V1 C2 V2 (C3).
- We only included the target vowels from previous articles: [i], [e], [a] and [o]. These vowels were equally represented in each vowel position, appearing 6 times as the V1 and 6 times as the V2. This provides two occurrences of each V1-V2 combination (a-o, a-e, a-i, o-a, o-e, o-i, e-a, e-o, e-i, and i-a, i-o, i-e).
- We included the target consonants from previous studies, [k], [p],[s] and [t] at the start of the word (C1 position) for at least 4 words, and in the middle of 3 words (C2 position).

In addition to these criteria, consonant clusters were avoided, as were diphthongs. To ensure that words were known by all participants, we controlled their frequency of apparition in an adult lexicon. The frequency of occurrence out of 1 million words averages at 55.31

(min = 0.26, max = 880.76, *SD* = 180.25). The completed word list (see Table 1) fulfills these criteria. Several other consonants that have not been analyzed previously were included in these words. Indeed, [b, d, f, ʃ, m] appear in the initial C1 position and [ʃ, n, l, m, g, ʁ, d, z] in the middle C2 position.

Table 1:

Whistled words chosen and tested

Word	IPA form//	Vowel int	Word	IPA form//	Vowel int
Bateau	bato	1	Béquille	bekij	1
Cassis	kasis	2	Cocher	koʃe	2
Copie	kopi	3	Chameau	ʃamo	1
Dépôt	depo	2	Finale	final	2
Fossé	fose	2	Kilo	Kilo	3
Mégot	mego	2	Peril	peʁil	1
Passé	pase	1	Petard	petaʁ	1
Piquet	pike	1	Police	polis	3
Sachet	saʃe	1	Sauna	sona	1
Sirop	siʁo	3	Soda	soda	1
Tapis	tapi	2	Têtard	tetaʁ	1
Ticket	tike	1	Tisane	tizan	2

Due to the equal distribution of the whistled vowel pairs, the distribution of vowel frequency intervals is as follows: twelve pairs are at a relative distance of 1, eight pairs are at a relative distance of 2, and four at a relative distance of 3. These intervals are clearly perceptible in the spectrograph of whistled vowels, where larger vowel intervals show a larger pitch movement, compared to smaller intervals (Figure 1)[11].

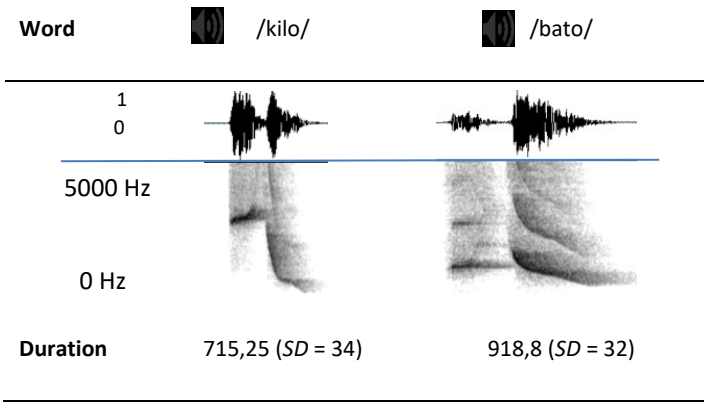
A single whistler provided us with these whistled word recordings, recorded on a Zoom H1 with the assistance of Julien Meyer (the third author). This whistler is fluent in the whistled

[11] For a full representation of the signal and spectrogram of the whistled words used, see Table 3 in Annex A.3. To compare the whistled word signal with modal speech, see Table 4 in Annex, A.3.

Spanish technique and also speaks French sufficiently well to follow French word prosody and to pronounce the vowels and consonants of the corpus as a French speaker would. The recordings consisted of a spoken version of the word (used to control the pronunciation) followed by the whistled version, which was repeated 4 times (note that due to over articulation of words in whistled speech, it was less necessary to introduce the whistled words in a carrier sentence, even if frequencies at the end of CVCV words still tend to lower more than if presented in a carrier sentence, especially for /o/ and /a/).

Figure 1:

Signal and spectrogram of /kilo/ and /bato/



Variability in whistled words of the corpus

In the whistled word recordings selected, there is a certain amount of variability. Though these differences are primarily due to the differences between words, there are also variations within the four productions of each word used in this experiment. The transformation from modal speech towards whistled words eliminates a number of cues present in the spoken versions, thus variation can be found within the duration, and in the transformations of the salient characteristics of each word, notably the whistled pitches and amplitude.

In terms of word duration, it is generally observed that whistled words are longer than spoken words (Meyer, 2015). The average whistled word in our corpus has a duration of 834 ms (*SD* = 110), compared to 530 ms in modal speech (*SD* = 130). The correlation between the two durations is not significant (Pearson's correlation r(22)= .37)[12]. The variability in duration between words is similar between the whistled words and the modal spoken words, though slightly lower for whistled speech[13]. Because the correlation is not statistically significant, we consider that word duration cannot be used for recognition in our task. We also note that for whistled words (with the exception of /ʃamo/), the duration of the second syllable is systematically longer than the first syllable[14], in agreement with French prosody of spoken words.

When considering the vowel pitches within the whistled words tested, we find some differences between pitches produced in the C1V1C2V2(C3) form according to position (see Figure 2). The vowel frequencies for /i,e,a,o/ in V1 and V2 positions remain within a similar range, corresponding to that of the vowels tested in previous studies (Meyer, 2015, Tran Ngoc et al., 2020b); however we found the V2 vowels to be much more stable than the V1 vowels for /a/, /e/ and /o/ (as seen in Figure 2). This is less applicable for /i/, which presents the most variability (this effect on /i/, which was also found for isolated vowels, Tran Ngoc et al., 2020b, may be due to the fact that its production requires higher efforts and constraints, particularly while whistling). Another difference in vowel production according to position applies to /o/, which is much higher in V1 than in V2 (an average frequency of 1453.3 Hz vs. 1137.9 Hz), and is therefore quite close to /a/ in V1. This effect corresponds to a largely observed tendency to lower the /o/ at the end of a speech utterance in Silbo (Busnel & Classe, 1976; Diaz, 2017;

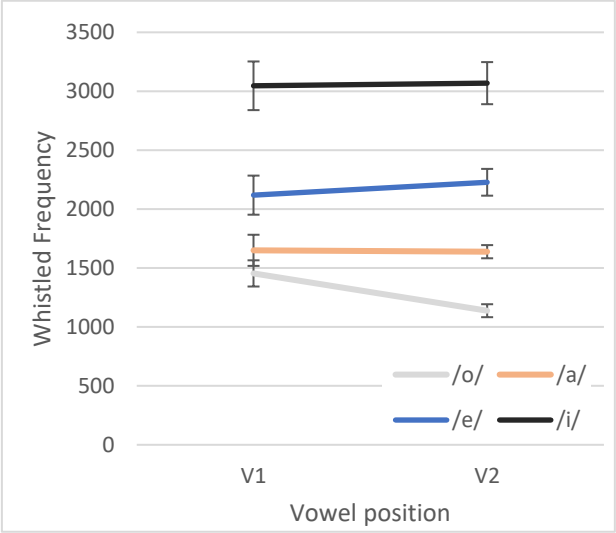
[12] See Annex, A.3, Figure 2
[13] See Annex A.3, Figure 3
[14] See Annex A.3, Figure 4

Meyer, 2015), nonetheless, we note the difference with the average frequency of the isolated /o/ at 1230 Hz (Tran Ngoc et al., 2020b).

Yet, the similarity in frequencies (between V1/o/ and V1/a/) is not necessarily problematic for recognition, as the vowel position is often based on the creation of a relative vowel space and distance between each of the vowels (see the relative distances proposed in Tran Ngoc et al., 2020b). We note that this stable relative distance may compensate for the variability of the pitches.

Figure 2:

Vowel frequency comparison between V1 and V2 positions in the words tested here



Design

The word options proposed to participants corresponded to 2 different lists of 5 possible answers per stimuli. We retained this forced-choice option given the novelty of this experiment, and the possibility that naïve participants may not succeed at recognizing words at all if an open choice option was given. To select the filler words, the list of 24 target words was randomized using https://www.random.org/lists/, and the first 4 options were selected (excluding the correct answer if present). This method was applied twice for every target word to construct two lists, A and B. This ensured randomness and variability when presenting the

answers. Because of this method, when considering the word options present for all 4 variations of the same word heard, certain word options can appear several times and others will never be proposed.

Procedure

Before starting the experiment, participants answered a short questionnaire, asking for their native language, age, gender and any musical experience. To present the format of the experiment to participants, a whistled example word was presented, /pate/ (“pâté”), along with a practice version of the answer format (a drop-down menu). Once they began the experiment, they heard a word in whistled speech selected randomly from the word list and were asked to pick the corresponding word from a list of five choices (which included the correct answer). Each of the 24 words are presented 4 times (four different recordings), for a total of 96 words heard. Once participants chose a word from the drop-down list, they were asked to validate their answer before moving on to the next word, giving participants the chance to think and change their mind before continuing. The next whistled word was played immediately after the validation of the previous answer. The possible answers appeared as soon as the participant clicked on the drop-down menu. Thus, participants first heard the word and then viewed the possible responses.

Participants

Nineteen participants were included here and gave informed consent before starting the experiment. They were all native French speakers aged between 18 and 36 years old (average = 25.57 years old; *SD* = 3.404). They had no language or hearing impairments and had no significant musical experience (as verified by a preliminary questionnaire). This group

included 12 women and 7 men. This experiment was conducted in accordance with the Helsinki agreement.

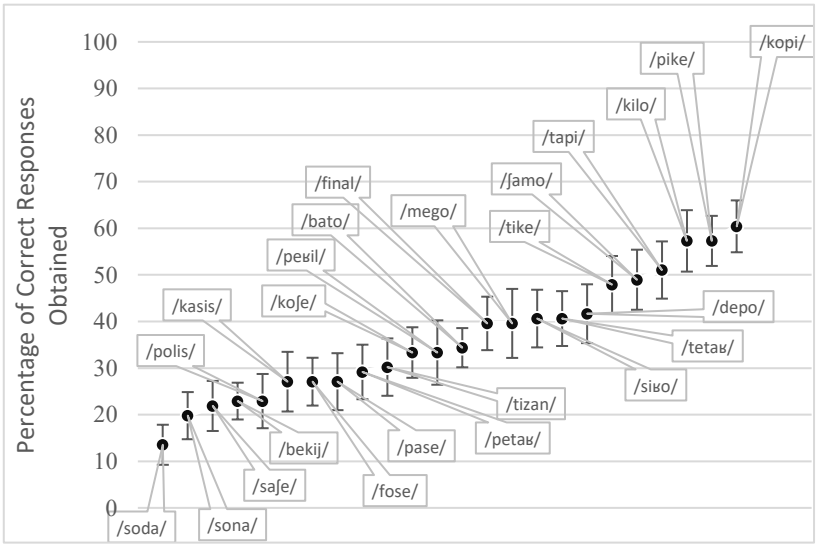
Results

Word Perception

We first considered overall word recognition results, with 96 responses for each of the participants, for a total of 1824 data points. Overall, we find that whistled words are recognized correctly with a rate of 45.6% of correct responses obtained. This value is well over chance, which is at 20% as there were five word options presented. However, the recognition rate varies greatly depending on the word played (see Figure 3): words like /soda/ and /sona/ are recognized under chance (at 13.5% and 19.29% respectively), and the words /bekij/, /saʃe/ and /polis/ just over chance (at around 22-23% of correct responses obtained). On the other side of the spectrum, words like /kilo/, /pike/ and /kopi/ are recognized much better: at 57.29% for /kilo/ and /pike/ and 60.41% for /kopi/. We notice that words containing the highest and lowest vowel frequencies (thus the most contrasting) have a higher average percentage of correct responses (more specifically for /kopi/ and /kilo/).

Figure 3:

Percentage of Correct responses obtained per word with standard error shown



Comparison between correspondence of Vowels heard and answered and Consonants heard and answered

In order to explore the role of vowels and consonants, we coded performances by targeting response rates for these elements for all of the correct and incorrect answers at the words level. We applied a Generalized Linear Mixed Model (GLMM) to explore how vowel matches (between vowels belonging to the word played and vowels included in the word answered) compare to consonant matches (consonants in the word played and consonants in the word answered). We included Phoneme (Consonant, Vowel) and Position (1, 2) as fixed factors, with Word and Participant as random effects. This showed a significant effect of Phoneme ($X^2(1, N = 19) = 97.25$, $p < .001$), of Position ($X^2(1, N = 19) = 6.44$, $p = .011$) and an interaction between Phoneme*Position ($X^2(1, N = 19) = 18.68$, $p < .001$). The application of a post-hoc test on this interaction (Bonferroni correction used for all post-hoc tests in this paper) demonstrated that V1 influenced participants' choices more than C1, that V2 influenced participants' choices more than C2, and also that V2 influenced participants' choices more than V1 (ps <.001). This suggested a difference in the role attributed to vowels and consonants, and an impact of vowel position.

Vowel position and the played/answered correspondence

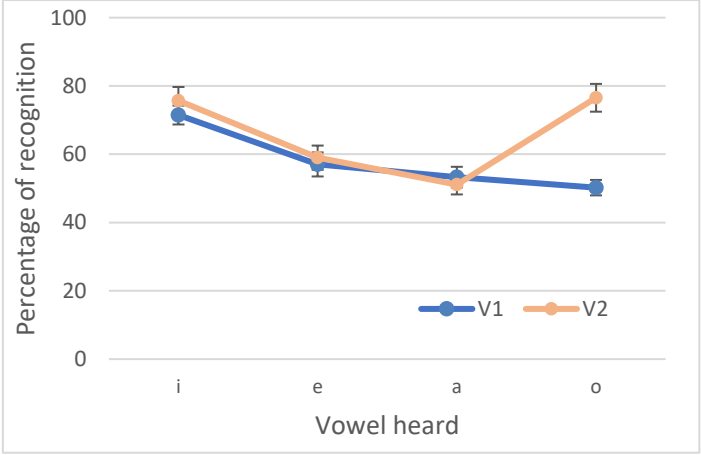
To explore the impact of vowel position, we measured the vowel correspondence within the word. Taking into account all of the answers given (correct or incorrect) at the word level, we applied a GLMM on the correspondence between vowels in the word played and vowels in the word answered. We included Vowel played (/i,e,a,o/) and Vowel position (1, 2) as fixed factors, and Participants as a random factor. We found a significant effect of Vowel played (X^2 (3, N = 19) = 99.0, $p < .001$), and a significant main effect of Vowel position (X^2 (1, N

= 19) = 24.9 *p* <.001). We also found a significant interaction between Vowel played*Vowel position (X^2 (3, N = 19) = 49.6, *p* <.001).

A post-hoc test revealed a significant difference between the two positions for the vowel /o/, where V1/o/ < V2/o/ (*p*<.001). We also find different relationships between the vowels according to position. For V1, we find that /i/ > /a/, /i/ > /e/, and /i/ > /o/ (*p*s<.001) giving us the hierarchy /i > o = a = e/, underlining that /i/ was recognized best. For V2, we find significant differences between /i/ and /o/ and the other vowels: /i/ > /a/ and /i/ > /e/ (*p*s<.001), and /o/ > /a/ and /o/ > /e/ (*p*s<.001). This gives us the following vowel hierarchy: /i = o > a = e/, underlining how /i/ and /o/ are recognized better than /a/ and /e/ (see Figure 4).

Figure 4:

Effect of Vowel position on Vowel recognition



Vowel interval

To investigate the effects of vowel interval on vowel recognition rates, we conducted a GLMM on Correct Answers with Vowel interval (1, 2 or 3) as a fixed factor and Participants as a random effect. We find a significant effect of Vowel interval (X^2 (1, N = 19) = 26.4, *p* <.001). Specific comparisons show significant differences between intervals, where 3 > 2 (*p* =.003) and 3 > 1 (*p* <.001). This shows a clear advantage for the largest interval, i.e. 3.

Discussion

In our analysis of whistled words, it is clear that, despite the strong phonetic reduction at play in whistled transformations, untrained participants recognize whistled words well over chance. However, while using an experimental paradigm that favored recognition (forced choice between 5 possibilities), the rate of word recognition observed here (45.6%) is far from the 20-30% increase in results observed for native whistlers when identifying whistled words in comparison to isolated phonemes (Meyer, 2015, Busnel, 1970; Tran Ngoc et al., 2020b; Tran Ngoc et al., 2022a). This suggests a difference in recognition strategies between naive listeners and native whistlers, possibly due to more active top-down processes of lexical access in trained listeners. We also found significant differences between certain words, helping to specify such processes.

In our analysis of the correspondence rates for vowels and consonants in the words answered, we observe a stronger influence of the vowels. Though this may be due to the higher diversity of consonants in the test, these results suggest that vowels serve a different role than consonants in whistled word recognition for naive listeners (where they are notably more affected by vowel position/context). Indeed, the effect of vowel position that we found suggested an impact of context on vowel contributions for word recognition in disyllabic words (in line with Delle Luche et al., 2014) for spoken words). In furthering our analysis of the vowel played and its position, we found that the position affects categorization hierarchy. In V1, words with /i/ are recognized better than words with any of the other vowel. However, in V2, this advantage is shared with words containing /o/, giving the same vowel hierarchy as found in previous studies with isolated vowels (/i>o>a=e/) (Tran Ngoc et al., 2020b). We suggest that this difference is due to the proximity between the frequencies of /o/ and /a/ in

V1. Another possible explanation could be the stability of V2 and the longer duration of the second syllable.

A more detailed exploration of the vocalic context in disyllabic words showed that larger vowel frequency differences coined better whistled word recognition for naive listeners (we found an effect of relative vowel frequency interval, where we demonstrate that the largest interval (n=3) is significantly better recognized than the smaller ones (n=1 and 2)). This confirms and reinforces the advantage found in previous studies for /i/ and /o/ - respectively the highest and lowest whistled vowels in Spanish - as they represent the boundaries of the largest frequency intervals and of the whistled vowel space. Interestingly, skilled whistlers and Silbo teachers (Diaz, 2017) also mention the importance of relative frequencies and use it to explain better vowel recognition performances.

Conclusions

In conclusion, naive listeners recognize whistled words correctly and well over chance using whistled vowel pitches, and frequency intervals. These results demonstrate a difference between the roles of vowels and consonants in whistled word recognition, as well as the importance of vowel context.

Chapter 4

Musical expertise and whistled phonemes

Introduction

The second objective of this thesis is to study the impact of listener variability, and more specifically, musical experience, on the perception of whistled speech. Having shown that naive (non-musician) listeners can categorize whistled vowels and consonants correctly and well over chance, we now investigate how musical experience will affect categorization rates and phoneme preferences. In our previous research, focusing on transfers between music and speech, we underlined the importance of similarities between the two, in order for transfer to take place (Chpt.1.3.1 and Chpt.1.3.2). Here, as we consider whistled speech, we can first compare the similarities between whistled speech and music.

Whistled speech and music

When comparing whistled speech and music, many of the similarities between modal speech and music are maintained, however, the transposition from modal speech into whistled speech also likens the signal to music in other ways. First, when we consider the pitch of whistled speech, the range of whistled frequencies of Silbo lies between 1000 and 4000 Hz (Meyer, 2015), much higher than modal speech for men and women (F0 between 85-155 Hz for men, and 165-255 Hz for women). This whistled range is much closer to that of musical instruments. The violin, for example, can play fundamental frequencies between 196 Hz and 3500 Hz (G3 to A7), the flute plays between 261 Hz and 2100 Hz (C4 and C7), and the piano

between 27.5 Hz and 4200 Hz (A0 to C8). Classical singers, whose highest fundamental frequency is still much lower than that of the whistled vowels, 880 Hz for an A5 being the highest note sung by a soprano (with the exclusion of whistled tone singing, which can reach much higher frequencies), also learn to shape their vocal tract and cluster formants in the 2500-3500 Hz range, thereby increasing the volume of their productions (Sundberg, 2003; Cohen, 2019). Thus, by using formants in a pitch range similar to those found in whistled speech (Meyer, 2015), singers can project their voice over the sound of an orchestra and into the audience. The pitch of musical instruments (including voice) is therefore more similar to the whistled pitch range than modal speech is. In addition, and in contrast to modal speech, a whistler-specific frequency range is established where relative intervals are maintained between the vowel pitches (see Chpt.3.1). This resembles a musical structure such as a scale or chord. However, the purpose of pitch in music and whistled speech remains fundamentally different. Whistled pitches reflect aspects of the spoken vowel timbre and not speech prosody, thus the whistled pitch is closely connected to meaning, while music is not (see Chpt.1.3.1.2).

When considering the musical timbre of whistled speech (in terms of its sound quality) it is also easy to hear a similarity with music. This is in part a consequence of how the sound is produced, i.e. the use of blown air. Indeed, whistled speech produces a sound with qualities such as bright, breathy, or clear. Such terms could easily describe other wind instruments, such as the flute (see Reymore, 2021), or even the oboe or clarinet. Many forms of “singing” whistling, as described by Meyer (2015) use such instruments (flutes, or whistles). These similarities suggest that musical experience would impact whistled speech perception, showing a transfer of knowledge between the two.

We therefore seek to answer the following question in both articles: How do participants with musical experience categorize whistled phonemes?

Each of these articles considers different aspects of musical experience using the whistled vowel and whistled consonant categorization tasks.

Articles

In the first article presented here, we consider whistled vowel categorization in a behavioral test (Expt 6) reprising the 3-part structure of experiments presented in previous chapters (Expts 3 & 4). Therefore, like these previous experiments, we also include inter- and intra-whistler variability and a training portion with feedback. After considering the effect of musical experience on whistled vowel categorization, as well as the impact of whistler, we investigate the way in which whistled speech is processed by participants with a musical background. This question is particularly relevant for the whistled vowel, as the vowels contain very few whistled cues and easily resemble music notes. Thus for the vowel to be categorized correctly, we suggest that it must be considered as speech rather than music.

The second article presented in this chapter focuses on whistled consonant categorization with a single behavioral experiment (Expt 7). This experiment is based on the previously presented whistled consonant categorization experiments in Chapter 2 (Expts 1A & 1B), and uses an identical 3-part structure, including a training portion with feedback, and the integration of intra-whistler variability in the 3rd part. In this article, we consider the effect of musical experience on whistled consonant categorization by addressing the idea of knowledge transfer, and taking an interest in the type of skills transferred from music to whistled speech. We wonder which of these skills produces a musical advantage. To approach this question, musical knowledge was considered in two different ways: according to general

music knowledge, and according to instrument specialization, thus including specific sounds and techniques used when playing this instrument. As the whistled speech experiments also included a comment section with specific questions, participants' answers seem particularly relevant to understanding how whistled speech was treated. These have been included in the Annex (A.4.1)

In these two articles, we regroup participants according to musical levels differently. Traditionally, musical experience has been categorized in a binary and often arbitrary way (see Chpt.1.3.3.1). Here, we asked participants to indicate their musical experience according to a scale ranging from 0 (no-musical experience) to 6 (professional musician), where level 4 corresponds to having obtained the DEM diploma (see Chpt.1.3.3.2). Thus, in our first approach towards musical experience, we chose to divide the participants with musical experience according to a cut-off based on musical diplomas, creating a group of "non-musicians" with participants whose musical levels were 0, 1, 2, and 3, and "musicians", participants with levels 4, 5 and 6. In many ways this differentiation is supported by previous studies and definitions of the "musician", though it may not be a realistic representation of skill transfer. Thus, in Experiment 7, we divided the participants into three groups: non-musicians (level 0), low-level musicians (levels 1, 2, and 3) and high-level musicians (levels 4, 5 and 6). Finally, by including instrument specialization, we take into account the neurological differences shown according to instrument timbre, and instrument-specific knowledge (as shown in Chpt.1.3.2.1).

An overview of this chapter and its experiments are presented in Table 6.

Table 6:

Description of the experiments in the articles of Chapter 4

Chapter 4 – Musicians	Articles	Expt	Target	Reference	Participants	Musical Groups 0 1 2 3 4 5 6
4.1	Tran Ngoc et al., 2023b	Expt 6	Musical experience & vowels	Expt 3/4	67	Non-mus / Mus
4.2	Tran Ngoc et al., 2023c	Expt 7	Musical experience & cons	Expt 1A/1B	66	None / Low / High / *Instr

Instrument specific knowledge

In choosing to consider musical instrument specialization, we sought to represent each instrument family: wind instruments (with the flute), string instruments (with the violin), percussion instruments (with the piano, though it is also a string instrument), and singing (or voice), because of its proximity with speech. Including participants with expert knowledge of these instruments can easily lead us to question the possibility of instrument-based predisposition. However, up until very recently, children were rarely oriented toward finding a fitting instrument for their interests, skills, and bodies. Indeed, in French music conservatories, there are often waiting lists for socially prized instruments such as the violin and piano (“Listes d’attentes: une fatalité”, 2015), due to a general lack of teacher availability (Pégourie, 2015). In Tranchant’s study (2016), these waiting lists also encourage families to be involved in the conservatory as early as possible to have a better chance of picking their instrument, and this choice is sometimes influenced by the child’s family. In addition, there have often been strong gender roles associated with the choice of instrument. Some instruments, like the brass sections, are usually male-dominant, whereas others, such as the flute and the harp, are female-dominant (Sergeant & Himonides, 2019). These instrument stereotypes and preferences are reflected in the instrument choices in music schools, where

there is a clear effect of social pressure (Montandon, 2018). Today, parents are generally encouraged to introduce different sounds to their child, to ask the child for their opinion, and to consider the propensity for certain instruments (Kubik, 2016), however, the underlying influence of social norms, and family pressure may eliminate one's choice of instrument, making instrument predisposition difficult to imagine, or to test.

Each instrument family (string, wind, percussion – thus excluding voice) uses specific techniques, materials, and production methods. The differences between these families are loosely based on the discrimination between solid-body vibrating instruments and air-vibrating instruments. This differentiation may also consider the materials used but rarely relies on the instrument timbre (Rognoni, 2008; Adler, 2002).

Wind instruments, or aerophones[15] can be divided into families according to the sound production methods, and "original" materials: woodwinds (made of wood, with or without single or double reeds) and brass. String instruments are chordophones[16], categorized according to the sound production method: "bowed" instruments (violin for example), plucked string instruments (harp), and hammered string instruments (piano for example). Percussive instruments produce sound when hit, thus the hammered string instruments (piano, clavichord, and dulcimer) can also be considered percussion instruments. Finally, voice is often excluded from these instrument groups (Cohen, 2019), because it is not an instrument in the physical sense (thus not included in Adler, 2002), even though there is a specific training of the vocal cords and voice mechanisms. By choosing to represent an instrument from each

[15] Aerophones are instruments that produce sound using air.
[16] Chordophones are instruments that use strings. Often the vibrating string will transfer energy to a solid structure

family in our study, we can compare production mechanisms (air, hammered, bowed), as well as various forms of musical articulation.

The violin creates an attack or musical articulation according to the weight put on the bow and the speed of the bow (Adler, 2002). The flute produces sound by blowing air over the embouchure hole and using phonetic articulation to produce attacks (Adler, 2002). In flute, a technique known as tonguing is used to articulate notes in wind instruments (Tuley, 2021), is usually constructed using a CV form such as /te/, and the articulations most frequently taught and discussed are /p/, /b/, /t/, /d/, /k/ and /g/. The piano, contrary to both the violin and the flute, is a polyphonic percussive instrument that produces sound through the striking of small, padded hammers onto the strings, thus reducing articulation to staccato or legato (see Bresin & Batel, 2000). The voice, or human's natural musical instrument, is both subtle and flexible, yet almost always linked to text, borrowing from speech for producing articulation. This diversity in instruments therefore provides sufficient differences in skill-sets, allowing us to gain insight into the musical listener's knowledge.

4.1 The Effect of musical expertise on whistled vowel identification

Abstract

In this paper, we looked at the impact of musical experience on whistled vowel categorization by native French speakers. Whistled speech, a natural, yet modified speech type, augments speech amplitude while transposing the signal to a range of fairly high frequencies, i.e. 1 to 4kHz. The whistled vowels are simple pitches of different heights depending on the vowel position, and generally represent the most stable part of the signal, just as in modal speech. They are modulated by consonant coarticulation(s), resulting in characteristic pitch movements. This change in speech mode can liken the speech signal to musical notes and their modulations; however, the mechanisms used to categorize whistled phonemes rely on abstract phonological knowledge and representation. Here we explore the impact of musical expertise on such a process by focusing on four whistled vowels (/i, e, a, o/) which have been used in previous experiments with non-musicians. We also included inter-speaker production variations, adding variability to the vowel pitches. Our results showed that all participants categorize whistled vowels well over chance, with musicians showing advantages for the middle whistled vowels (/a/ and /e/) as well as for the lower whistled vowel /o/. The whistler variability also affects musicians more than non-musicians and impacts their advantage, notably for the vowels /e/ and /o/. However, we find no specific learning advantage for musicians, but rather learning effects for /a/ and /e/ when taking into account all participants. This suggests that though musical experience may help structure the vowel hierarchy when the whistler has a larger range, this advantage cannot be generalized when listening to another whistler. Thus, the transfer of musical knowledge present in this task only influences certain aspects of speech perception.

Article Information

Article Status

This article is under review in the journal Speech and Communication:

Tran Ngoc, A., Meyer, J. & Meunier, F. (2023b). The Effect of Musical Expertise on Whistled Vowel Identification. *Speech and Communication*.

Key Words: vowel categorization, whistled speech, acoustic cues, musicians, musical experience, speech perception

The Effect of Musical Expertise on Whistled Vowel Identification

Introduction

Whistled speech is a form of naturally modified speech that transposes spoken (modal) speech into whistles (see Meyer, 2015, for a review). Though whistled speech is intelligible only to trained speakers, previous studies have shown that naive listeners can categorize whistled phonemes correctly and better than chance (see for example Meyer et al., 2017). In most non-tonal languages, the transposition from modal speech to whistled speech relies on a 'formant-based whistling strategy' (Leroy, 1970; Busnel & CLasse, 1976; Rialland, 2005; Meyer, 2015), where whistlers make an approximation of the vocal tract articulation in the spoken form to produce the whistled form. This translates to different whistled pitch ranges emitted for each vowel type. This inter-vocalic pitch difference is a simplified whistled reflection of the different spoken frequency distributions characterizing distinct vowels. Vocalic whistled frequencies can be related to specific formant distributions or more generally to different spoken vocalic timbres. In whistled Spanish for example, /i/ has the highest mean pitch value, /e/ is slightly lower, /a/ is even lower and /o/ has the lowest mean frequency, while /u/ is quite low and generally overlaps strongly with /o/ (Meyer, 2008; Diaz, 2008). The intra-vocalic pitch variation observed within the range of whistled frequencies covered by each vowel type depends on several factors including consonant coarticulation, stress, and/or position in the word. There are also some inter-individual variations due to different whistling techniques, different communication distances or to differences in the size of the front oral cavity, specific to each individual (Busnel & CLasse, 1976; Diaz, 2008; Meyer, 2015).

Previous behavioral studies on whistled speech have analyzed the four Spanish vowels /i, e, a, o/ extensively, in conditions which verified that the vowel frequency ranges tested did not overlap. They have demonstrated how naive French listeners are able to categorize these whistled vowels correctly well above chance (Meyer, 2005; Meyer et al., 2017; Tran Ngoc et al., 2020b). In addition, vowels demonstrated the same categorization hierarchy across studies, despite having been extracted from different whistlers and different types of contexts (words vs. sentences, long distance vs. middle distance whistles). Overall, for naive French listeners, /i/ was always recognized best (with 78% to 86% correct answers, depending on the study), followed by /o/ (50 to 59% correct answers), /e/ (44 to 47%), and finally /a/ (38 to 44%). Previous papers have attributed this hierarchy to the frequency distribution of the vowels: one argument was that /e/ and /a/, vowels at intermediate frequency ranges, have two frequency neighbors, while /i/ and /o/ have only one. Another argument has been made based on formant convergence in spoken speech, which also explained the advantage for /i/ (Meyer et al., 2017). The 2017 study integrated listeners with different native languages and showed the impact of listener experience on vowel recognition, where the vowel categorization rate was modulated according to one's native language (Spanish, French or Standard Chinese). The results of the native French and Spanish speakers were not significantly different for correct answers, reflecting the very close vocalic characteristics across languages of the four whistled vowels tested /i, e, a, o/, but they also showed differences in confusions (Meyer, 2008; Meyer et al., 2017), revealing the influence of the different vowel system of each listener's native language. More recently, two studies have focused on whistler variability (Tran Ngoc et al., 2020b; Tran Ngoc et al., 2023e) by including two different whistlers. They once again confirmed the vowel hierarchy previously observed.

Here, we build on these previous studies, keeping the variability in productions, and go further by looking at how listeners' musical experience affects vowel categorization, given that the whistled speech form may resemble an instrumental mode more than a spoken mode at first listen.

Musical Experience

Musical experience is shown to have a positive impact on speech perception in a variety of conditions and tasks. This includes improved phonological discrimination compared to non-musicians when listening to L2 sounds (Slevc & Miyake, 2006) or to speech in noisy environments, like in the street or a crowded room (Bidelman & Krishnan, 2010; Strait & Kraus, 2011, Varnet et al., 2015). Such advantages also extend to the discrimination of vowel and consonant sounds (Parbery-Clark et al., 2012), where musicians treat voiced and unvoiced stimuli differently than non-musicians (Ott et al., 2011). Differences in whistled speech perception according to musical experience have also been suggested by Meyer (2008), as the few participants with musical experience included in his whistled vowel perception test showed improved results compared to non-musicians. In addition, another recent study on whistled consonants showed an effect of musical experience and specific instrument training on consonant categorization (Tran Ngoc et al., 2022b).

As musical training involves learning to identify and distinguish different pitches, rhythms, and tones (elements that are also present in speech), it would appear that these skills can be transferred to perceiving and processing speech sounds. Among musicians, pitch perception is developed along two different axes, both relevant to speech perception. The first is the auditory perception of frequency, where musicians have been shown to distinguish pitch changes more accurately than non-musicians (Tervaniemi et al., 2005) and with lower

frequency discrimination thresholds (Liang et al., 2016). The second is the ability to categorize pitch, encouraged by ear training, often giving rise to abilities such as absolute pitch (giving the name of the note heard, Ross et al., 2005). Such skills have been shown to apply to speech perception in noise (Varnet et al., 2015), and to tone-language perception by non-tone language speakers (Han et al., 2019), where musicians show clear advantages over non-musicians.

Peretz and Coltheart (2003) suggest a mechanistic model for sound processing, where sounds are treated as either music or speech, starting with a common "acoustic analysis" which then feeds-forward either to a music-specific module ("contour analysis"), a language-specific module ("acoustic-to-phonological conversion"), or to a still-not-characterized module ("rhythm and meter"). The initial shared perceptive capacities therefore explain a possible crossover between the two perceptive systems. Moreover, like speech, musical experience also engages various cognitive processes such as attention, memory or executive functions processes (Schellenberg & Weiss, 2013; Parbery-Clark et al., 2009).

Whistled speech blurs the boundaries between the cues typically used to distinguish speech and those used for music, as the whistled vowel pitches reflect aspects of the speaker's timbre. Thus, though musicians may identify the whistled pitches and pitch movement more easily due to improved frequency discrimination, to categorize the vowels correctly, musician participants must integrate such information as part of their phonological representation of the vowels. Swaminathan and Schellenberg (2017), who demonstrated that rhythm training in music did not affect consonant categorization in Zulu clicks for English speakers, suggest that musical competence is only relevant for meaningful cues (see Kraus & Chandrasekaran, 2010). Therefore, the whistled vowels would need to be heard as speech (rather than music) to correctly identify them despite these reduced cues, using what Peretz and Coltheart called

the “acoustic-to-phonological conversion”. Because of these constraints, we wanted to explore if, and how, musical experience serves as an advantage for whistled vowel categorization.

Inter-whistler variability

In this study, we integrated inter-whistler variability by using stimuli from two whistlers with different vowel spaces, building on two previous experiments with inter-whistler variability (Tran Ngoc et al., 2020a; Tran Ngoc et al., 2023e), which tested only non-musician participants. While initial studies on whistled speech have included some intra-talker variability, few studies addressed inter-talker variability in whistled speech, despite research showing that inter-talker variability has significant effects on spoken speech perception presented in adverse conditions (Zaar & Dau, 2015). Indeed, Zaar & Dau (2015) show that the largest perceptual variability was induced by across-talker variability for the same CV items, with different CV confusions according to the speaker. Correlations between certain acoustic phonetic properties and listener comprehension have also been observed for usual modal spoken speech with non-native listeners (Bent et al., 2010), where talkers with a larger vowel space were easier to understand. An experiment displaying a combination of these conditions (native and non-native listeners with inter-talker variability and presented in noise) showed results in the same line, with a significant effect of inter-talker variability on intelligibility (Dommelen & Hazen, 2012). The acoustic phonetic properties which trigger an effect on speech perception, though they depend on whether listeners are native or non-native, include more energy in the 1-3 kHz range and an enlarged vowel space in the F2 range. Interestingly, the stimuli from these previous experiments on modal speech deal with constraints which also characterize whistled speech (such as the 1 to 4kHz range of whistled pitch and the importance

of the F2/F3 range in spoken-to-whistled transpositions, see Meyer, 2015) allowing us to investigate the impact of acoustic phonetic inter-talker variations in whistled speech perception. Results from our previous studies testing non-musicians (Tran Ngoc et al., 2020b; Tran Ngoc et al., 2023e) showed an advantage for the whistler with a larger range, whose whistled vowel productions were easier to categorize, demonstrating an impact of inter-whistler variability.

Here, we add to these conditions, by considering the impact of musical experience on whistled vowel categorization in addition to the inter- and intra- variabilities within the productions of two whistlers. We also explore the possibility of a learning effect throughout the different parts of the experiment. To do so, we used a three-part experiment construction, like in one of our previous experiments (Tran Ngoc et al., 2020b). Part 1 asks participants to respond to stimuli without any previous introduction, part 2 proposes a very short learning phase where feedback is given, and finally part 3 consists of the same test as part 1, but with stimuli from the other whistler. In one previous study non-musician participants heard the same whistler throughout the experiment (Tran Ngoc et al., 2023e) and we did not observe an overall learning effect. However, there was a significant improvement of correct categorization rates appearing only for the whistler with the most restricted range and only for the vowel /e/. In the experiment presented here, we included two different whistlers in each list presented to participants, while maintaining the possibility of testing the effect of the whistled vowel range as the different whistlers were presented separately (one in each part). Thus, we took into consideration a possible effect of inter-talker variability, all while questioning the musician's ability to adapt to an individual whistler-specific frequency distribution.

Experiment

Method

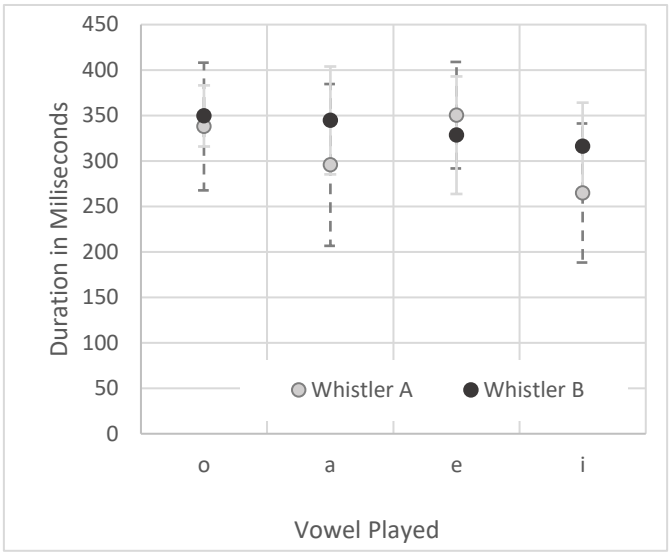
Stimuli

The second author recorded the stimuli in a soundproof room of the Gipsa-Lab (with a H4N-Zoom audio recorder) using the built-in stereo microphone placed 2 m from the whistler to avoid near acoustic effects of the source. Two different expert whistlers were recorded, both teachers of whistled speech in the Canary Islands, and a stereo-to-mono conversion was applied to the recordings (using Matlab). The whistled Spanish vowels targeted, /i/, /e/, /a/ and /o/ were extracted from disyllabic CVCV whistled words. To maintain a similar prosody for each stimulus vowel, we selected vowels from the second unaccented CV syllable. Moreover, to guarantee a realistic variability in terms of coarticulation contexts, we selected vowels following various consonant attacks (/d/, /k/, /g/, /t/), where, after removing the consonant attack (only the vocalic nuclei were kept), silence was added to the vowels to create homogenous samples of 500 milliseconds. We selected 48 stimuli per whistler (96 in total) from these recordings, corresponding to 12 versions of each vowel, where 3 different recordings were extracted from the same consonantal context. Due to this limited number of items, the objective was not to test the influence of such contexts, but rather to maintain some aspects of variability due to coarticulation between consonantal and vocalic segments in the experiment. The duration of these 96 whistled vowels also varied, lasting between 146 ms to 473 ms (both by whistler A). The variance for whistler A (6.36 ms) is slightly higher than for whistler B (2.78 ms). When comparing the average duration of each vowel, it appears that /i/ is slightly shorter for both whistlers, as it is in modal speech, though these durations are variable for each of the vowels produced (see Figure 1). An ANOVA with repeated measures

on vowel duration, with Whistler and Vowel type as factors, shows a significant difference between Whistlers ($F_{1,11}$ = 6.60, p =.026), with whistler B's productions being significantly longer than those of whistler A, and a significant difference between Vowels ($F_{3,33}$ = 3.75, p =.02), but no Whistlers*Vowels interaction. The application of post-hoc tests for specific comparisons with Bonferroni corrections only shows /o/ to be significantly longer than /i/ (p =.03).

***Figure 1**:*

Distribution of vowel duration average with presentation of the standard deviation according to whistler.

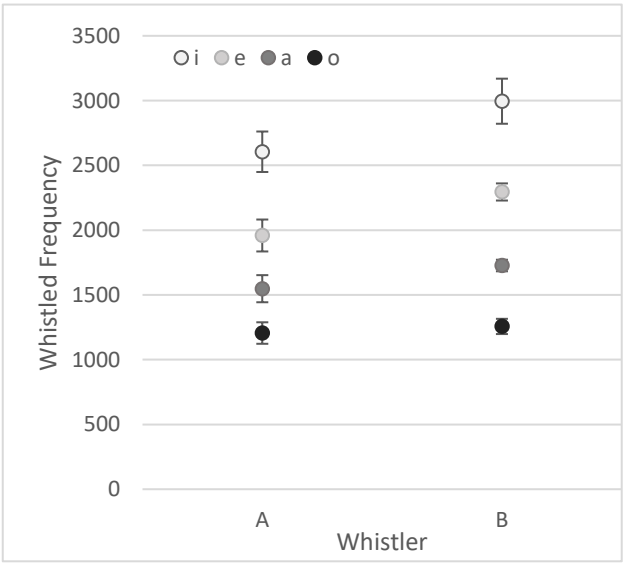


In terms of frequency, the value attributed to each vowel corresponded to the whistled vowel frequency average measured across the duration of each vowel nucleus (thus excluding rapid frequency modulations of consonant attacks). Certain vowel groups vary more than others (see Figure 2). This is partly due to the fact that some vowels require more strength for whistled production (especially the front /i/ which is most acute, followed by /e/), and to the varying consonantal contexts from which the vowels were extracted, even though we retained only the part of the vowel containing a stable frequency. Typically, coarticulation with coronal consonants push vowel frequencies to higher values, whereas velar consonants pull towards

lower values (Meyer et al., 2019; Tran Ngoc et al., 2022a). This is especially the case for high frequency whistled vowels, /i/ and /e/ (SD = 256.64 Hz, SD = 196.99 Hz respectively), where we note that /i/ frequencies vary strongly for both whistlers, whereas /e/ variations are more present for whistler A than for whistler B. By contrast, /a/ and /o/ are more stable for both whistlers (SD = 120.6 Hz and SD = 75 Hz respectively). Although there is some variability for each vowel, the main difference between the recordings occurs in the vowel range of the whistlers. Indeed, the vowel frequencies of whistler A are proportionately less spread out than those of whistler B, whereas the vowel groups of whistler B are more distanced from one another and the frequencies are more stable, with the exception of /i/ (see Figure 2).

Figure 2:

Variability in mean frequencies of whistled vowel productions for each whistler and each vowel



Design

The design used is identical to the one in Tran Ngoc et al., 2020b. In the first part of the experiment, we evaluated how naive participants performed on whistled vowel categorization when hearing one whistler's productions. In this part, we randomly presented 48 stimuli to the participants, corresponding to 12 versions of each vowel type. These include 3 different

recordings of each vowel, extracted from the same consonantal context. After this first part, a training section with feedback ensued, comprising 16 vowels with 4 recordings of each vowel (each corresponding to a different consonant attack), which used the vowel productions of the same whistler as part 1. In a third part, participants listen to the stimuli from the other whistler, which consist of 48 whistled vowels (12 versions of each vowel type, with the same criteria as part 1). This design enabled us to test whether participants with musical experience showed an overall learning effect between parts, or whether listeners rely on a single relative pitch scale. Overall two lists, both containing three parts, i.e. part 1 (test), part 2 (training) and part 3 (test), were tested: one with whistler A first and one with whistler B first. Each participant was presented with only one of the lists. The experiment was proposed online and was programmed with PCIbex Farm. Thus, participation took place at home using headphones, earbuds or speakers.

Procedure

Before starting the experiment, participants were asked to fill out a questionnaire aiming to collect information on their language and musical backgrounds. We asked participants to indicate their age, the languages they speak and the level of each language (rated on a three-level scale: “beginner”, “intermediate”, “confirmed”). They were also asked to indicate their musical experience, including the instrument played, the level achieved for each instrument, and their background in said instrument (location or context of lessons and how many years of experience they had). As an indication of each participant’s instrument level, we asked participants to choose between 1 – beginner (*Débutant*), 2 – amateur, 3 – confirmed (*Confirmé*), 4 – DEM musical Diploma (*DEM*), 5 – Superior University Diploma (*Diplôme Supérieur*) and 6 – professional musician (*Professionnel*). These levels were chosen specifically to target French classical musicians by including music diplomas such as the DEM.

After the questionnaire was completed, the experiment itself started. Part 1 presents participants with recordings performed by one of the whistlers. It asks participants to categorize the whistled vowels heard without any training, and using the arrow keys. The arrow keys are attributed to each vowel following the keyboard layout, and are visually presented before and during the experiment. After giving participants instructions explaining the task, they are presented with the stimuli in a random order. Each recording is played once, and responses are accepted 500 ms after the start of the recording, to ensure that participants listened to the entire clip before responding. The next recording is played 200 ms after the participant's response. In the second part (part 2), participants then complete a short training session with feedback for four versions of each whistled vowel (from the same whistler as the one heard in part 1). These stimuli are presented in a random order. The feedback, which was shown as soon as participants responded, consisted of either "No, this was not the correct response" (*Non ce n'était pas la bonne réponse*), or "Congratulations!" (*Bravo!*). The feedback was shown for 1000 ms, before moving on to the following vowel. All participants were given the same form of training regardless of their musical experience. Finally, in part 3, participants are asked to categorize the whistled vowels of the other whistler (if they heard whistler A in parts 1 and 2 now they will hear whistler B and the reverse if they first heard whistler B). Aside from using the other whistler's recordings, this part is identical to part 1.

Participants

This experiment was conducted in accordance with the Helsinki agreement. Sixty-seven participants were included in this study. They were native French speakers who had no language impairments nor any previous knowledge of whistled speech. They were all between 18 and 50 years old (M = 27.25, SD = 7.16).

We chose to divide the participants into two groups according to the levels of musical expertise, opposing those with a high-level of musical skill (as verified through diplomas), which we called “musicians” (levels 4, 5 and 6) and the low-level musicians or participants without musical experience, which we called “non-musicians” (levels 0, 1, 2 and 3). This echoes previous definitions of the “musician”, as a form of knowledge that is skill based, and generally tested in a performance-related capacity. Indeed, citing Hallam (2010), as described by Zhang et al. (2020), we propose that a musician is “someone who has the ability to play a musical instrument”. We therefore defined participants’ musical level according to participants’ main instrument level (often declared first and defined as the instrument on which they have attained the highest level).

The non-musician group included 30 participants, with 22 women and 8 men, who had an average age of 29.5 years old (*SD* = 8.79). In this group, 10 participants had no musical experience whatsoever; however, the non-musician group also included 4 participants who were beginners, 7 amateurs and 9 participants with confirmed musical experience. The musician group included 37 musician participants with 18 women and 19 men, with an average age of 25.37 (*SD* = 4.89). This group contained 18 musicians who had obtained their DEM diploma, 8 musicians with a Superior Diploma and 11 professional musicians.

All of the participants recruited spoke a second language, with a majority speaking English. As the whistled phonemes were based on a Spanish form of whistling and produced by Spanish speakers, we took a special interest in participants who have experience with Spanish. In total 35 out of 67 participants spoke Spanish, where 17 participants had a “beginner” level, 14 participants had an “intermediate” level, and only 4 had a “confirmed” level. Thirty-two participants spoke no Spanish.

Overall, 37 participants heard whistler A first, and 30 participants heard whistler B first. When considering our two musical experience-based groups: among the non-musicians, 15 participants heard whistler A first and 15 participants heard whistler B first; In the musician group, 22 participants heard whistler A first, and 15 heard whistler B first.

Results

In our vowel analysis, we considered the answers for 48 items in part 1 and for 48 items in part 3 for a total of 96 items for each participant. Two musician participants were excluded as they performed outside 2 standard deviations from their group. We therefore considered the data of 65 participants (30 non-musicians and 35 musicians) for a total of 6240 items in this analysis. We find that overall, participants categorized the whistled vowels correctly with 60.83% of correct responses obtained (*SD* = 13.5), well over chance at 25%.

Before running our main analyses and as we used Spanish vowel productions (though we underline the similarity between French vowels and the Spanish vowels chosen), we performed an ANOVA on Correct Answers to test whether experience with this language influenced overall vowel categorization rates. We included Spanish Level as a variable. This revealed no significant effect ($F < 1$).

We first ran a Generalized Linear Mixed Model (GLMM) on Correct Answers. We included four fixed factors: Musical Experience (musician/non-musician), Part (P1, P3), Whistler (A, B) and Vowel Type (/a/, /e/, /i/, /o/). We included Participant as a random factor.

There is a significant main effect of Vowel (X^2 (3, N=65) = 435.381, p <.001). It appears that /i/ was categorized best (at 80.8%), followed by /o/ (at 63.9%), both of which were much better categorized than /a/ and /e/ (at 50.8% and 47.8% respectively). Post-hoc tests (with

Bonferroni corrections applied throughout the results section) revealed that all vowels give significantly different performances from each other (*p*s <.001) except for /a/ and /e/. We also observed a significant main effect of Whistler, X^2 (1, *N*=65) = 65.3, *p* <.001, where 65.3% of correct answers (*SD* = 17.3) are obtained for whistler B and 56.3% (*SD* = 13.1) for whistler A. There was also a significant effect of Musical Experience (X^2(1, *N*=65) = 4.772, *p* =.029), where musicians obtained 64.6% (*SD* = 13.7) of correct responses, and non-musicians obtained 56.4% (*SD* = 12) of correct responses. There were no significant effects for Part (*p* >.05) suggesting that there was, overall, no general learning effect.

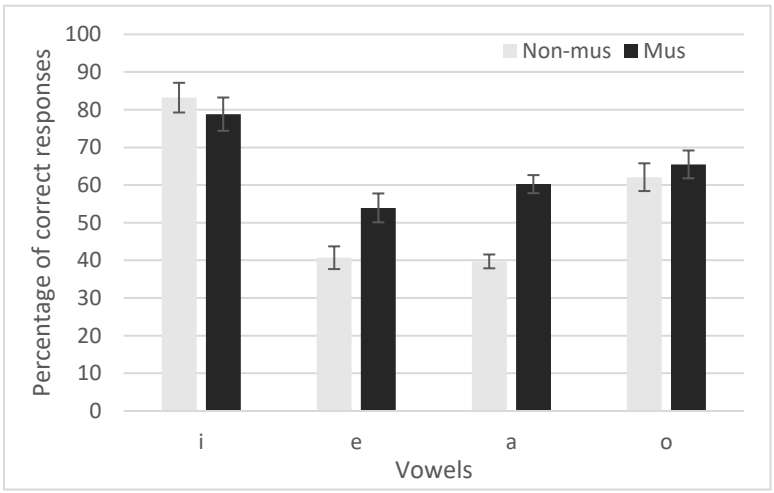
We observed a significant interaction between Musical Experience and Whistler (X^2 (3, *N*=65) = 6.357, *p* =.012). Post-hoc tests reveal that while the difference between the groups is not significant for whistler A, there is a significant difference between musicians and non-musicians for whistler B (71.1% compared to 58.5%, *p* =.023). The differences between the whistlers are significant for both groups but bigger for the musician group. Indeed, for the musician group, whistler B gives rise to performances that were 13% higher (*p* <.001, where M = 71.13%, *SD* = 16.1 for B; and M = 58.09%, *SD* = 14.15 for A), while for the non-musician group, performances for whistler B were only 4 % higher (*p* =.001, M = 54.31%, M = 58.54%, *SD* = 16.44 for B, *SD* = 11.73 for A).

We also observed three simple interactions in which the factor Vowel type interacted significantly: with Whistler (X^2 (3, *N*=65) = 21.363, *p* <.001), Part (X^2 (3, *N*=65) = 25.310, *p* <.001), and Musical Experience (X^2 (3, *N*=65) = 60.495, *p* <.001). These interactions suggest that the effect of each of these factors depends on the vowel played. Moreover, two double interactions reached significance: Musical Experience*Whistler*Vowel (X^2 (3, *N*=65) = 38.776, *p* <.001) and Whistler*Part*Vowel (X^2 (3, *N*=65) = 15.124, *p* =.002). In order to understand

these results and interactions, we analyzed the data independently for each vowel (see Figure 3).

Figure 3:

Average percentage of correct responses for musicians and non-musician participants, with standard error indicated for each vowel.



For each vowel, we applied a GLMM with Musical Experience (musician, non-musician), Part (P1, P3), and Whistler (A, B) as fixed factors. We included Participant as a random effect. We note that for each of these analyses *df* = 1 and N = 65. All post-hoc analyses used Bonferroni corrections.

For the vowel /i/ there is a significant main effect of Musical Experience (X^2 = 4.49, *p* =.034) showing a difference between the results of non-musicians and musicians, where non-musicians (M = 83.12%, *SD* = 10.85) perform better than musician participants (M = 78.8%, *SD* = 13.15). There is a significant effect of Whistler (X^2 = 36.3386, *p* <.001), where performances with whistler B give rise to 86.41% of correct categorization, and those with whistler A give rise to 75.25% of correct responses. Thus, the results obtained for whistler B are superior to those obtained by whistler A. However, the significant interaction between Musical Experience*Whistler (*p* =.008) reveals that this difference is significant only for non-musician participants (*p* <.001) with 91.4% of correct answers for whistler B and 75.5% of correct

answers for whistler A. We also observed a difference between groups only for whistler B, where musicians obtained only 82.14% of correct answers.

For /e/, there is a significant effect of Musical Experience (X^2 = 8.1567, *p* =.004) which shows that musician participants (M = 53.92%, *SD* = 21) perform better than non-musicians (M = 40.69%, *SD* = 17.84). We also find a significant effect of Part (X^2 = 10.37, *p* =.001) for which performances are higher in part 3 (52.4%) than in part 1 (43%). For this vowel, one simple interaction is significant between Musical Experience*Whistler, (X^2 = 26.0908, *p* <.001) where we find a significant difference between whistlers A (45%) and B (62.9%) only for musicians (*p* <.001). We also find a significant difference between musicians and non-musicians, though only within the results of whistler B (*p* <.001), with 62.9% of correct responses for musicians, and 36.7% of correct responses for non-musicians.

For the vowel /a/, we find a main significant effect of Musical Experience (X^2 = 12.6636, *p* <.001), which shows that musicians (M = 60.23%, *SD* = 24.24) categorize this vowel better than non-musicians do (M = 39.72%, *SD* = 23.30). We also observe a significant effect between parts (X^2 = 13.5121, *p* <.001), for which performances are higher in part 3 (54.4%) than in part 1 (42.4%), and a significant effect of Whistler (X^2 = 4.6066, *p* =.032), where, when hearing whistler A, participants obtain 47.9% of correct responses, compared to 53.56% for whistler B. No interaction reaches significance.

We observed two significant main effects for the vowel /o/, for Whistler (X^2 = 34.482, *p* <.001), where responses for whistler B are at 70.5% and responses for Whistler A are at 57.31%, and for Part (X^2 = 6.608, *p* =.01), where participants obtain 65.6% in part 1 and 62.18% in part 3. We observe a significant interaction between Musical Experience*Whistler (X^2 = 19.638, *p* <.001), where the application of a post-hoc test shows a significant difference between

whistlers for musician participants (p <.001), where whistler B responses are at 76.2% and whistler A responses at 54.8%. We also find a significant difference between musicians (76.2%) and non-musicians (63.8%) for whistler B. We recapitulated the significant differences per vowel in Table 1.

Table 1:

Comparison of post-hoc test results per vowel, only significant effects are shown

Whistled Vowel		**/i/**	**/e/**	**/a/**	**/o/**
Musical Experience		Mus < Non-mus	Mus > Non-mus	Mus > Non-mus	
Whistler		A < B		A < B	A < B
Part			P1 < P3	P1 < P3	P1 > P3
*Mus*Whist*	*Musicians*		A < B		A < B
	Non-musicians	A < B			
	Whist B	Mus < Non- Mus	Mus > Non-mus		Mus > Non-mus

Discussion

In this experiment, we looked at how whistled vowels are categorized by naive listeners depending on their musical expertise. We aimed to extend previous results and look at how musical expertise modulates performance when faced with whistler variability and a short training segment with feedback. First, the data confirmed previous results from other experiments on non-musicians. Overall, whistled vowel categorization was obtained with an average of 60.8% of correct responses (well over chance at 25%), and we replicated the hierarchy between the vowel categorization performances: /i/ is better recognized than /o/, which is better recognized that /a/ and /e/ (which were not different from each other), leading to the hierarchy i > o > a = e.

Concerning the role of musical expertise, our results are in line with the literature, showing better performances for musicians than for non-musicians. However, the interactions showed that this advantage is in fact more specific than it may seem, as different effects are observed depending on the vowel produced and the whistler, specifying the role of musical experience. Interestingly enough, the effect of musical expertise is clear for the least well-recognized vowels (/e/, /a/, and /o/), where for /e/ and /o/ this advantage is specific to the productions of the whistler with the widest range of frequencies (whistler B). For /a/ the advantage for productions of whistler B over A is general, applying to both musicians and non-musicians. Overall, this suggests that musical expertise mainly affects the categorization of the whistler with the larger vowel frequency range. This effect appears for vowels /e/, /a/ and /o/ that show less pitch variability across different productions. These same vowels are also more stable in whistler B's productions than in whistler A's.

However, for /i/, the best-recognized vowel, the pattern of answers is different. Indeed, only non-musicians show a preference for whistler B, for whom they also show better results than the musician participants do (9.26% difference). This could be due to a higher variability in the whistled frequencies of /i/ (Figure 2), which may affect musicians more strongly. As musicians are better than non-musicians at distinguishing pitch differences (see Tervaniemi et al., 2005, Liang et al., 2016) this non-musician advantage may be due to an over-reliance on pitch matching, or over-sensitivity to these variations. By contrast, non-musicians' ability to better assimilate the whistled /i/ to modal speech may partly come from more receptivity to secondary acoustic characteristics of vowels, such as duration. Indeed, in our corpus, /i/ is the vowel with the shortest duration, which matches with general spoken intrinsic durations observed in phonetics (House & Fairbanks, 1953; Solé & Ohala, 2010).

To summarize the inter-talker variability effects highlighted above, listening to whistler B, with a larger vowel frequency space than whistler A, leads to better performances than listening to whistler A. This effect exists for every vowel but at different levels: first, for /i/, we find a significant difference for non-musicians only (15.9% in favor of whistler B). For the two vowels /e/ and /o/, we only observed a significant difference for musicians (with respectively 17.9% and 12.4% better results on productions of whistler B), whereas performances for /a/ are significantly higher for all participants (5.66% in favor of whistler B).

Finally, we found no overall training effect; however, there was a significant increase in performances during the experiment for two specific vowels, notably for the less well-categorized vowels (/a/ by 12%, and /e/ by 9.4%). Interestingly, these training effects were not specific to musicians and were present regardless of the whistler heard. This complements results we highlighted earlier for /e/ and /a/, notably the fact that a musical advantage was found only on whistler B for these two vowels. This reinforces the interpretation that a finer auditory sensitivity to pitch in musicians is partly responsible for the result patterns. These last results on 'training' are partly in line with those of another study (Tran Ngoc et al., 2023e), which included only non-musician participants: findings showed no evidence of a global training effect, even when the same whistler was presented throughout the experiment. However, a significant improvement in correct categorization was observed for the vowel /e/ and only for the whistler with the most restricted range. This suggests that participants quickly improve their recognition of the least well-categorized vowels. Intriguingly, in the present experiment we also observe a slight decrease between parts 1 and 3 (-3.4%) for /o/, the second best categorized vowel. We wonder whether having whistler A (with a smaller range) in the first part of the experiment may be disturbing, leading listeners to perform worse than during their initial perception. However, given the large difference between the whistlers and

the lack of simple interaction between the whistlers and the parts, it would be inappropriate to elaborate on these dynamic effects and further experiments should be conducted to explore this point.

Nonetheless, and in line with existing literature (Dommelen & Hazen, 2012), our findings suggest that more stable frequencies and a larger vowel space facilitate abstract representations of certain vowels. Previous studies have clearly shown that the task proposed in this experiment triggers whistled vowel categorization with different answer patterns. These have been observed between populations of different language backgrounds tested on the same protocol, but with productions of a different whistler. These findings reflects the influence of different vowel spaces from the listeners' mother tongues (Meyer et al., 2017). As our present task requires the listeners to perform a pitch-to-timbre matching/association between the whistled vocalic pitch heard and the mentally recalled vowel quality, this process is eager to be influenced by acoustic parameters other than pitch. Some examples include parameters defining the formants of each vowel, as well as formant proximities, already known to be important for vowel identity (see Meyer et al., 2017, for a discussion). As musicians in occidental musical traditions are trained to recognize pitches and interpret them as notes (pitch categorization), we wonder whether the stability of the vowels with the lower whistled frequencies, as well as the length of whistler B's pitches could then allow musicians to better construct the relative relationships between pitches (especially for /e/ and /o/). Musician participants may also exploit these cues more efficiently due to their enhanced auditory sensitivity. The higher variability found in /i/ in terms of frequency might hinder such skills, partly explaining the non-musician advantage for that vowel, as they may also rely on different cues than musician participants. This suggests that musicians process the whistled signal differently than non-musicians. However, the musicians' advantages are often specific

to whistler B. Thus, though we can consider that musicians are able to use their phonological knowledge to categorize whistled vowels well over chance, they exploit the musical similarities in timbre or frequency discrimination only when the signal is stable and thus possibly more similar to musical notes.

Conclusion

In conclusion, naive French listeners with and without a high level of musical experience recognize whistled vowels more than 50 % of the time. These results appear to be robust and generalizable. Having a high-level of musical experience proves to be beneficial for the vowels whistled with lower frequencies (/e,a,o/), though this is specific to one of the two whistlers for /e,o/. This advantage was not uniform, and depended on the stimuli according to specific acoustic conditions. We evidence the impact of the whistler heard, where the whistler with the larger range was categorized better than the other whistler, which is apparent for both musician participants and non-musician participants, though more often specific to musician participants. This musical advantage, observed mainly for one whistler, may be due to the stability of the whistled pitches, allowing for better exploitation of pitch-based skills (such as relative interval definition). It also further underlines the influence of a whistler's range and frequency distribution on vowel categorization, showing that a larger vowel space facilitates the creation of abstract vowel representations for both musicians and non-musicians.

4.2 Benefits of musical experience on whistled consonant categorization: analyzing the cognitive transfer processes

Abstract

In this study, we explored the transfer of musical knowledge and skills towards speech perception, by analyzing the perception and categorization of consonants pronounced in a naturally modified speech form known as whistled speech. This speech mode, used for long distance communication, is characterized by a simple modulated melodic line, akin to a musical tonal line. We conducted this study with two aims: (i) to explore the effects of different levels of musical experience on speech perception, and (ii) to better understand the type of knowledge transferred, by focusing on the group of participants with a high-level of musical skill. Within this high-level group, we opposed the multidimensional cognitive transfer with sound-specific transfers by considering instrument specialization, and opposing general musical knowledge (common to all instruments) with instrument-specific training. We focused on four instruments: voice, violin, piano and flute. Our results confirm the presence of a general musical advantage, and suggest that only a small amount of musical experience is necessary for a transfer of skills towards whistled speech perception. However, a higher level of skill achieved has a stronger effect, specifically for certain consonants. Our results also show that instrument expertise has an effect on whistled speech perception. Thus the transfer of knowledge cannot be attributed solely to general musical experience, affecting general cognitive functions such as executive functions, memory or attention, but rather our findings show that the modification in whistled speech processing is essentially due to specific acoustic familiarization (possibly linked to production) in high-level musicians.

Article Information

Article Status

This article has been submitted to Glossa Psycholinguistics.

Tran Ngoc, A., Meyer, J. & Meunier, F. (2023c). Benefits of musical experience on whistled consonant categorization. *Glossa Psycholinguistics*.

Keywords: consonant categorization, musicians, perception, whistled speech

Data Accessibility Statement: The data from this experiment is available here.

Benefits of musical experience on whistled consonant categorization

Introduction

Musical experience and whistled speech

Music and speech are similar in many ways. From an acoustic perspective, both signals are complex, containing a melody, rhythm, syntax and smaller units. From a cognitive perspective, both music and speech share various cognitive processes such as attention, memory or executive functions, each implicated when storing sounds and structure, and activated through production/perception. These similarities have led some to consider that knowledge can be transferred from music to language, and this has been demonstrated on various levels of speech processing, including phonological awareness (Bhide et al., 2013), learning new words (Barbaroux, 2019) or perceiving speech in noise (Straight & Kraus, 2011). Indeed, changes in the brain through musical training provide a starting point for establishing such transfers, including modifications in the development of the temporal and frontal areas (Gaser & Schlaug, 2003), and an increase in grey matter or in cortical thickness (more specifically in the somatosensory cortex, linked to physical contact with the instrument, Bermudez et al., 2009). These changes are linked to differences in behavior: modifications in the temporal lobe (which includes the auditory cortex), for example, can correspond to improved sound perception capacities, and increased grey matter in the frontal areas can be linked to executive functions. Such capacities are thought to create a "fine-tuned" auditory system (Straight & Kraus, 2011; Smit et al., 2023), establishing advantages in speech-based tasks for "musicians". However, as pointed out by Smit et al. (2023), empirical evidence of these differences is not straightforward, as the specificities of the individual musician's brain

are bypassed and the definition of the musician is usually binary, established somewhat arbitrarily and based on variable factors, making it difficult to distinctly measure such changes. This further impedes our understanding of musical transfers towards speech processing, as musical knowledge and skill-sets are often poorly defined in experimental studies, and rarely take into account the multidimensional aspects involved (such as cultural settings). In addition, these difficulties refrain studies from establishing the types of musical skills that affect transfers the most, as well as the elements of speech perception that are affected.

To further analyze the effect of musical experience on speech perception, we turned towards a form of naturally modified speech that deforms yet simplifies the signal: whistled speech. To the untrained ear, this speech form, characterized by a "melody" of whistled pitches, may sound like musical notes, thus lying within the blurred boundary between speech and music (Smit et al., 2023) and making it a perfect tool for exploring transfers between music and speech. Indeed, whistled speech, used by populations around the world living in mountainous regions or in dense forests to communicate at a distance, reduces the complexity of the spoken signal into a whistled form with a modulated frequency and amplitude. This melodic line, situated between 0.8 and 4 kHz, covers a band of frequencies much higher than those carried by one's voice. This line adapts to each language: for non-tonal languages like French and Spanish, whistled speech encodes the timbre of vowels –or vowel quality- into different pitches, and transposes some of the acoustic cues available in the spectral formants of spoken modal speech. Typically, /i/ is whistled with the highest mean values, /o/ the lowest, and /e/ and /a/ are intermediate (Meyer, 2015). Like in modal speech, consonants modify the vowels' stable frequencies through articulatory movements (see Figure 1), modulating and/or interrupting the pitches. Yet, despite these acoustic differences, the whistled speech mode is largely similar to non-modified modal speech, as it uses the

phonological characteristics present in the base language. These are expressed phonetically by tentatively transposing the detailed pronunciation of modal spoken speech. Therefore, whistled speech allows for the same duality of patterning, and the production of the same vocabulary as the spoken mode. This speech form exists in a number of different languages, as well as different dialects (this is the case in the Canary Islands, Meyer, 2015), and the similarities make whistled speech an ideal basis not only for studying transfers, but also for speech processing. This is because it highlights subtle effects (also present in adverse listening situations) all while maintaining a natural speech signal. We will use this speech form to better explore one's capacity for phonological categorization, an essential component in the speech perception process.

Several previous experiments have demonstrated that naive participants, who have never heard whistled speech previously, are able to categorize whistled vowels and consonants well above chance (e.g. Tran Ngoc et al., 2020b, Tran Ngoc et al., 2022a). These results show that even though participants had never been confronted with this speech form before, they were able to categorize the signal by using processes generally applied to modal speech. They reflect important oppositions between certain vowels in terms of opening and placement (front/back), as well as between consonants, where we noted a hierarchy showing that /s/ and /t/ are better categorized than /k/ and /p/ (i.e., /s = t > k = p/), thus favoring whistled speech cues with pitch changes (Tran Ngoc et al., 2022a). The different role of these cues has been further underlined in the comparison of consonant confusions, which determined a hierarchy of cues. Indeed, though consonants with whistled pitch changes (/s/ and /t/) were easily distinguished from consonants without pitch changes (/k/ and /p/), there was some confusion between /p/ (slower amplitude rise) and /s/. This could be explained by the articulation of these consonants (i.e. the initial amplitude modulation) or by the number

of common cues (see Tran Ngoc et al., 2022a). Studies on whistled speech have also highlighted the effects of individual experience, showing that participants' performances vary according to their native language (Meyer et al., 2017). Here, by focusing on musical training and skill, we take an interest in individual experience, and we explore how musical experience affects the categorization of whistled consonants. More specifically, we look at the impact of such experiments on the consonant hierarchies found in previous studies, where musical experience was not taken into account and the level of musical expertise was rather low (Tran Ngoc et al., 2020a, Tran Ngoc et al., 2022a).

Defining the musician

One of the difficulties with understanding musical transfer, as underlined by Smit et al. (2023), is the lack of homogeneity in measuring musical experience. Often, studies underline a binary opposition between musicians and non-musicians (which could be argued as a loose, but culturally defined concept in western classical music tradition). However, there is also a lack of consensus on the criteria used to define these groups. Indeed, the definition of the "musician" often corresponds to a minimum number of years of musical experience, i.e. six years of formal training (Zhang et al., 2020; Smit et al., 2023). However, other measurements are also used, including the starting age, musical diplomas, or the score on musical "tests" (AMMA- Advanced Measures of Music Audition, the Wing Test, or the Goldsmith Musicality Index, among others). In some meta-analyses studying far transfers stemming from music (Sala & Gobet, 2020; Cooper, 2020), the "musician" participants may not even follow such criteria, and often have very little musical experience (a maximum of 507 hours, or a little more than 2 years). In addition, there is little or no indication of the format or type of musical learning undergone (see Sala & Gobet, 2020; Bigand & Tillman, 2022), despite the fact that

the specific format and organization of musical training may optimize adaptability and enhance executive functions (Degé, 2021), allowing for a transfer of skills. Such disparity makes it additionally difficult to validate musical transfers to speech.

Although the previous experimental work used varied and inconsistent analytical criteria to define musicians (years of training, age, tests), these do reflect certain steps and components necessary for learning music. Yet because this process is quite complex, it is difficult to consider the full spectrum of skills acquired. In conservatories or music schools, such skills are tested through performance-based exams on a musician's instrument, where the act of performance requires and encompasses many individual skills (i.e. auditory, motor and cognitive skills), yet almost entirely disregards criteria such as the number of years of training or starting age. These exams then give authenticity to a musician's level and status. Though this system is culturally specific and generally associated with Western classical music, in countries like France this type of training is highly institutionalized, and instrumental skills are developed according to a defined scale. When targeting "musical experience" here, we propose that by using this well-established structure, we can better understand the various skill-levels of participants as defined in the French classical music world.

In specifying these skills, we underline that the "musician" is often in reality an instrumentalist (as demonstrated through performance-based exams), and that, the longer a musician trains, the time spent practicing their instrument surpasses that of general musical training (ear-training, rhythm, note-reading etc.). This instrument specialization can be observed on a neuro-functional level, as musicians show specific cerebral activation when listening to their own instrument (Pantev & Herholz, 2011; Pantev et al., 2001; Margulis et al., 2009). However, as suggested by the binary opposition between musicians and non-musicians which is generally applied, musical instrument specialization criteria is rarely used to nuance

and understand musical experience. Here, we chose to take this factor into consideration to explore the multidimensional musical knowledge in addition to the "level" of musical skill achieved, looking at how consonant categorization differs according to the level of musical expertise and instrument specialization.

Transferring perceptual and cognitive skills

Despite the difficulties with defining "musician" participants, the musical advantages shown for speech are numerous and occur on various levels (see Besson et al., 2011). On an initial perceptive level, studies show that there is a modification of the auditory threshold of musicians compared to non-musicians (Tervaniemi et al., 2005). This has led to a more general assumption that musical training helps to process sounds on an auditory level (Straight & Kraus, 2011; Kraus & Chandrasekaran, 2010; Smit et al., 2023, Varnet et al., 2015). Yet, on a more general cognitive level, one can also observe the transfer of musically learned abilities on short-term memory (Tierney et al., 2008), and on executive functions. According to some studies, these skills can even have an impact on one's performance on intelligence tests (Degé et al. 2011). However, the very possibility of a far-transfer is still being debated (Sala & Gobet, 2020 and Bigand & Tillman, 2022). As speech perception also implicates both auditory perception (when receiving the signal) and cognitive functions (to understand the elements heard), we wanted to explore how such skills could help understand musical advantages in speech perception. This reprises Barbaroux (2019)'s question, who opposes two hypotheses when describing the advantages of musical experience on speech perception. On one hand lies the "multidimensional" approach, where the advantages observed can be attributed to a number of cognitive functions improved by musical experience, and on the other hand, the "waterfall' approach, where the benefits observed would spring essentially from improved

auditory perception. Barbaroux's results, based on an experiment requiring participants to learn and categorize new words, support the "waterfall" hypothesis. In our experiment, we explore these two propositions, with the idea that instrument specific effects can be interpreted as supporting a "waterfall" interpretation.

To explore these issues, we ran an experiment based on Tran Ngoc, Meyer, Meunier, 2020a's study on whistled consonants with French speakers who had no particular musical experience. In the present paper, we included a total of 66 participants with approximately 10 semi-professional or professional classical musicians for each of 4 target instruments: the violin, the piano, the flute and voice. By choosing to integrate musicians with a classical music background in France (French speakers, who indicated having studies in a music conservatory), we consider that there is a homogeneity in musical knowledge among participants (both culturally and according to institutionalized examinations), as well as a clear differentiation in musical skills in terms of instrumental specialization. The four target instruments included in this study incorporate each of the different instrument families (string, percussion, wind and voice), with different timbres and production mechanisms. We can therefore take an interest not only in the differences in speech perception according to musical experience, but also instrument specialization. This will allow us to explore certain processes, which may explain the benefits of musical experience, and the type of transfer which occurs: if the effects of musical experience are essentially due to training general cognitive functions (i.e. the "multidimensional" model), instrumentalists should obtain similar results. However, if the performances of the instrumentalists vary according to the musical instrument played, advantages can be attributed to more specific instrument-based modifications affecting the perception of the signal (i.e. the "waterfall" model).

The categorization task proposed in our experiment requires participants to use various perceptual and cognitive skills. Firstly, auditory perception is essential to encode the differences between each of the consonants and to be able to associate the characteristics heard with one's own consonantal representation. This experiment also requires participants to use their memory to recall the sounds of the previous stimulus and their answer (notably during a training portion which incorporated feedback), as well as attentional skills, as the experiment is repetitive and requires focus. These elements will allow us to understand the transfer effect between perceptual and cognitive skills used in musical expertise and this categorization task more clearly.

Experiment

Method

Stimuli

In this experiment, we studied the four consonants [p], [t], [s], [k], three occlusive consonants ([p]-bilabial), [t]-dental/alveolar, [k]-velar) and a fricative consonant ([s]-alveolar). They were presented to listeners in their whistled form and are noted here using the phonemic transcription - typically appearing between slashes "/" - to represent these sounds. This choice enables us to use the same symbols to represent both the spoken reference in the mind of the listeners and the altered transformations into whistles with which they are confronted. The second author (Julien Meyer) recorded the stimuli in a soundproof room of the Gipsa-Lab (H4N Zoom recorder, using the built-in stereo microphone). Sound extracts were then converted to mono. A whistler knowledgeable in Silbo (the Spanish whistled language used in the Canary Islands) produced the whistled stimuli. Despite the difference in language, the whistled Spanish consonants used here were chosen because of their similarity

with French, allowing for a good categorization rate (well above that of chance, as demonstrated by Tran Ngoc et al., 2022a). These consonants were produced and presented in the VCV form /aCa/, to reduce the variations (due to co-articulation between vowels and consonants) and to involve the highest number of consonant specific cues. These stimuli include oppositions that allow for characteristic distinctions between whistled forms of consonants: 'acute/grave' and 'interrupted/continuous'. Among the selected consonants, we find /s/ to be acute and continuous, /t/ to be acute and interrupted, /k/ to be grave and interrupted and /p/ to be grave and interrupted (with a more gradual rise in amplitude, see Figure 1 in Tran Ngoc et al., 2020a).

The 16 recordings used as stimuli consist of 4 versions of /aka/, /apa/, /asa/ and /ata/. These recordings maintain a very consistent duration, with an average total length of 996 milliseconds (ms) (*SD* = 84.37). The first syllable (before the interruption of pitch, or the descent for /asa/) has an average duration of 324 ms (*SD* = 48.35) and the second syllable, an average duration of 671 ms (*SD* = 54.25). Although all the VCVs have very similar lengths, the variability in duration is slightly more important for /asa/ (where *SD* = 100 ms). It is also clear that the second 'syllable' (or CV) is longer than the first one (initial V), with an average difference of 227 ms (*SD* = 64.11). Using the program Praat, we calculated the average frequency of each of the vowels, before and after the consonant modulations. We find that, overall, the average frequency of the vowel /a/ is 1722.78 Hz, with little variation for the different consonants and productions (*SD* = 69.37). The frequency of the vowels preceding and following each consonant are also very consistent, deviating with an average of 48.9 Hz, *SD* = 28.36, with the final vowel being slightly higher than the first vowel (81% of the time). The consistency in stimuli duration and vowel frequency allows us to consider that any categorization differences between consonants should be attributed to specific consonant

cues. In addition to the main distinguishing cues, we notice small differences between consonants, notably between the maximum frequencies of the /s/ and /t/ productions (measured using Praat), where /asa/ reaches systematically higher frequencies than /ata/.

Design

The design used here is identical to Tran Ngoc et al.'s (2020a). The experiment consisted of 3 parts. In the first part of the experiment, we evaluated how naive participants (without any previous experience in whistled speech) performed on whistled consonant categorization. In this part, we randomly presented 40 stimuli to the participants, corresponding to one production of each consonant played 10 times each. Following this first part, a training section with feedback took place, comprised of 16 consonants, using each of the 4 recordings played 4 times each. In part 3, we tested participants' capacity for consonant categorization a second time, however, we included more variability in the stimuli. In this last part, the stimuli heard corresponded to 4 productions of each the syllables produced. Thus, in part 3, participants heard 12 different recordings (3 additional versions of each consonant were added to those heard in part 1), where each recording was played 3 times, giving a total of 48 recordings presented.

The experiment, programmed with PCIbex Farm, was proposed online. Therefore, participation took place at home using headphones, earbuds or speakers.

Procedure

Before starting this experiment, we asked participants to fill out a questionnaire indicating their musical experience, including the instrument played, the level achieved for each instrument and their background in that instrument (this included the number of years of experience or the context in which they took lessons - music conservatory, music school …).

We asked participants to choose between 1 – beginner (*Débutant*), 2 – amateur, 3 – confirmed (*Confirmé*), 4 – DEM musical Diploma (*DEM*), 5 – Superior University Diploma (*Diplôme Supérieur*) and 6 – professional musician (*Professionnel*). These levels were chosen specifically to target French classical musicians and instrumentalists. We relied on the rigor and organization of French music conservatories to distinguish classical musicians with instrument-specific diplomas (such as the DEM or the *Diplôme Supérieur*), from participants who did not have a strong instrument specialization, nor the common instruction required to obtain DEM or Superior University diplomas.

Then, participants heard a recording of each of the four whistled consonants without any indication of which consonant was played, to familiarize themselves with the sound quality of whistled speech (and thus its difference with the timbre of other instruments). We also presented the keyboard-based answer layout to participants before the start of the experiment.

Part 1 presents 40 recordings and participants are asked to categorize the whistled consonants without any training, using the arrow keys which are attributed to each consonant using the keyboard layout (shown on screen during the categorization tasks). In part 2, participants complete a short training session where feedback was given (either "Good Job" – *Bravo*, or "No this was not the correct response" – *Non ce n'était pas la bonne réponse*) after each categorization. Finally, in part 3, participants are once again asked to categorize 48 whistled consonant stimuli using the arrow keys. All these stimuli were presented in a random order.

Participants

This experiment includes 66 participants with 40 women and 26 men, with an average age of 27.2 years old (*SD* =7.03). All of the participants were native French speakers, with no

language problems, nor any previous knowledge of whistled speech. 36 participants (out of the 66) have a strong musical background, including flutists, pianists, violinists and singers who had achieved an « End of music school diploma » (level 4) at the very least. Among all 66 participants, 10 had no musical experience whatsoever, 4 participants were beginners, 7 amateurs, 9 with confirmed musical experience, 17 participants declared having level 4 (DEM diploma), 8 participants declared having achieved level 5 (Superior Diploma) and 11 were professional musicians (level 6).

Results

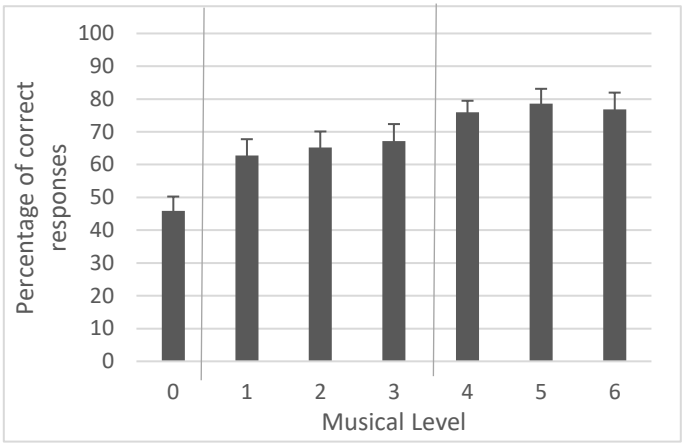
Differentiating musical transfer according to level of musical experience

We analyzed the data from parts 1 and 3, excluding the results from part 2 (with feedback), which included too few responses. Therefore, we analyzed 88 responses from each of the 66 participants, for a total of 5808 data components. The participants generally achieved an average of 69.1% (*SD* = 15.6) of correct responses, with chance at 25%, demonstrating a high level of whistled consonant categorization. In exploring the percentage of correct responses obtained according to the level (L) of musical skill, it appears clearly that the overall consonant performance rate increases according to the musical level obtained (except for the highest level- the professional level 6): L0 – 47,4%, L1 – 62,8%, L2 – 66,7%, L3 – 68,9%, L4 – 77,9%, L5 – 78,4%, L6 – 76% (see Figure 1). These differences show certain gaps between levels. In particular, two large gaps are found between Levels 0 and 1 (15.4% gap) and between Levels 3 and 4 (9% gap). Thus participants with no musical experience (level 0) can be differentiated from participants with a little musical experience (levels 1-3), and participants with a little musical experience (levels 1-3) can be differentiated from participants

with a high-level of musical experience (levels 4-6, confirmed musicians). Therefore, in the following analyses we separated our factor Musical experience into 3 groups of participants (no experience – level 0, low-level experience – levels 1-3, and high-level of experience – levels 4-6).

Figure 1:

Percentage of correct consonant responses according to Level



We ran a GLMM on Correct Answers (0, 1) with Consonant played (/k/, /p/, /s/, /t/), Part (P1, P3), and Musical Experience (None, Low, High) as fixed factors. We considered Participant to be a random effect. We find significant main effects for Musical Experience, Consonant, and Part. We also find three significant interactions: Part*Consonant, Consonant*Musical Experience and Part*Musical Experience. There were no significant double interactions. All post-hoc analyses are conducted using the Bonferroni correction.

There is a significant effect of Musical Experience (X^2 (2, *N*=66) = 31.55, $p < .001$), for which the post hoc test shows that participants with a high-level of musical experience (High), with 77.5% of correct responses (*SD* = 14.38), obtain better results than participants with a low-level of musical experience (Low), with 65.6% (*SD* = 13.12). In turn, participants with a low-level of musical experience perform better than participants with no musical experience

(None) with 45.9% (SD = 13.74). There is also a main effect of Part (X^2(1, *N*=66) = 6.01, *p* = .014), which shows that performances are higher in part 3 than in part 1. Consonant played also reached significance (X^2 (3, *N*=66) = 289.71, *p* <.001), where globally /s > t > k = p/.

We find a significant interaction between Part*Consonant, (X^2 (3, *N*=66) = 11.89, *p* =.008), which, through the post-hoc test, describes the evolution of individual consonant recognition. Specific comparison reveals that the learning effect between part 1 and part 3 is significant only for the consonant /t/, where participants obtain 69.4% of correct responses in part 1 and 79.4% in part 3 (*p* =.004). This learning effect is reflected in the hierarchies found in each part: in part 1, /s > t = k > p/ (*p*s <.05) and in part 3, /s = t > k > p/ (*p*s <.001).

The significant interaction between Part*Musical Experience (X^2 (2, *N*=67) = 20.56, *p* <.001) gives some insight into the effect of musical experience. With the application of a post-hoc test, we observe that participants with a high-level of musical experience (High) performed significantly better than participants with no musical experience (None) both in parts 1 (73.5% compared to 49.2%, *p* <.001) and 3 (80.7% compared to 43.1%, *p* =.004). In part 3, this difference according to musical experience is further nuanced, as we also observe a significant difference between participants with a low-level of musical experience (at 68.5%) and those without any experience (at 43.1%, *p* =.004), as well as between participants with a high-level of musical experience and those with a low-level (*p* =.026). The only significant learning effect (P1 < P3) observed is for the group of participants with a high-level of musical experience (*p* <.001).

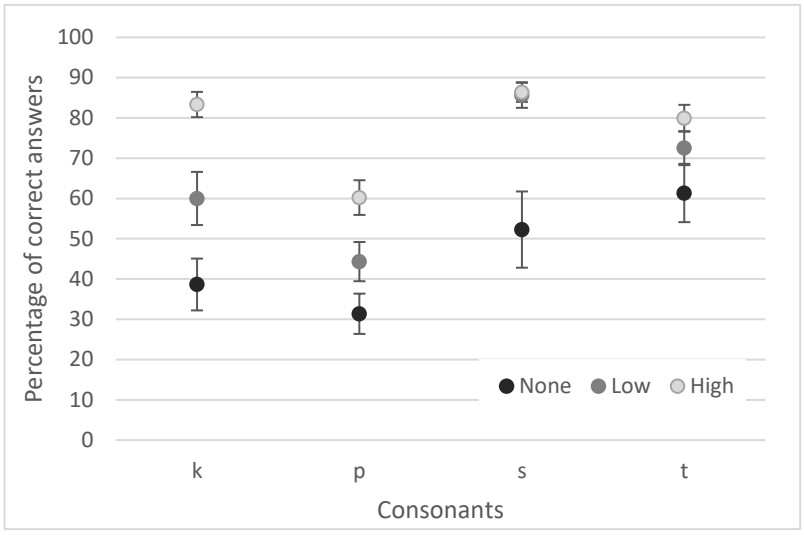
The interaction Consonant*Musical Experience is also significant, (X^2 (6, *N*=66) = 62.62, *p* <.001), and we observe significant differences between groups only for 3 consonants: /k/, /p/ and /s/. For /k/, we find significant differences between High (83.3%, *SD* = 18.85), Low

(60%, *SD* = 20.95), and None (38.63%, *SD* = 20.35), where participants with high-level musical experience perform significantly better than the other two groups (*p*s <.001). For /p/ we observe a significant difference only between High (60.22%, *SD* = 25.83) and None (31.36%, *SD* = 21.7), *p* =.009. Finally for /s/ we observe significant differences between High (86.36%, *SD* = 14.25) and None (52.27%, *SD* = 29.9), *p* <.001, and between Low (85.68%, *SD* = 14.16) and None, *p* <.001, where participants with musical experience (High and Low) have very similar (and high) categorization rates. This shows that though a high-level of musical experience shows advantages for 3 of the 4 consonants, even a low-level of musical experience has an effect, differentiating their results from those of participants with no musical experience. This applies especially to /s/ (see Figure 2).

These differences also reveal different consonant hierarchies between the groups. Participants with a high-level of musical experience show significant differences where /s > t > p/ (*p*s < .05) and /k > p/ (*p* < .001), thus giving the following hierarchy: /s (=k) > t (=k) > p/. The participants with a low-level of musical experience show significant differences where /s > t > k > p/ (*p* <.01). Finally participants with no musical experience show that /t > k > p/ and that /s > p/ (*p*s <.001), giving the following hierarchy: /t (=s) > k (=s) > p. This underlines the shift in participants' perception of /s/ according to musical experience.

Figure 2:

Percentage of correct responses obtained per consonant for participants with None, Low and High-Levels of musical experience



Thus, when considering the effect of musical experience on whistled consonant categorization, we observe significant differences between participants with a high-level of musical expertise compared to those with little or no musical experience (in both parts, with a learning effect for /t/, and specific advantages for /k/ and /p/). Thus, in further understanding the kind of transfer that occurs here, we reconsidered these high-level musicians through the instrument played.

Instrument specialization

To understand how musical experience differs according to the instrument played, we targeted 4 instruments: the violin (9 participants), the piano (7 participants), the flute (8 participants) and voice (7 participants). We excluded other high-level musician participants who did not play these target instruments, thus reducing the number of high-level musician participants to 31. We included all 30 participants with a low-level (Level 1 to 3) or no musical

experience (Level 0), therefore amounting to a total of 61 participants in this analysis, with 5368 data components. Though participants with a low-level of musical experience often play an instrument, we consider that their instrumental skill level may not be sufficient for instrument-specific differences. Therefore, we maintained low-level musicians as a separated group. We applied a GLMM to Correct Answers (0,1) with Instrument specialization (None, Low, Flute, Violin, Piano, Voice), Part (P1, P3) and Consonant (/s/, /k/, /p/, /t/) as fixed factors, with Participants as a random effect. Once again, all post-hoc tests were conducted with the Bonferroni correction.

We find significant main effects of Instrument (X^2 (5, *N*=61) = 51.94, *p* <.001), Part (X^2 (1, *N*=61) = 7.04, *p* =.008) and Consonant (X^2 (3, *N*=61) = 263.15, *p* <.001). We also find significant interactions for Part*Consonant (X^2 (3, *N*=61) = 12.26, *p* =.007), Instrument*Consonant (X^2 (15, *N*=61) =78.07, *p* <.001), Instrument*Part (X^2 (5, *N*=61) = 19.72, *p* =.001, and a double interaction Instrument*Part*Consonant (X^2 (15, *N*=61) = 24.98, *p* =.050). As we have already analyzed most of these effects previously in the comparison of the different levels, we will focus on the effects that include the factor Instrument.

The descriptive analyses of the Instrument showed that the highest average score is obtained by flutists (87.3% of correct answers, *SD* = 10.29), followed by singers (76.6%, *SD* = 13.17), violinists (75%, *SD* = 16.86), and pianists (74.3%, *SD* = 12.24); while participants with a low-level of musical experience obtained 65.6% (*SD* = 27.48), and participants with no musical experience obtained 45.9% (SD = 26.26). When applying post-hoc tests, we observe significant differences between participants with no musical experience and every instrument: Flute > None (*p* <.001), Violin > None (*p* <.001), Piano > None (*p* =.004) and Voice > None (*p* <.001), in addition to Low > None (p = .019) as found previously. We also observe a significant difference

between flutists and participants with a low-level of musical experience (Flute > Low, p <.001), as well as between flutists and pianists (Flute > Piano, p =.01). Overall, these results suggest that flutists categorized consonants best.

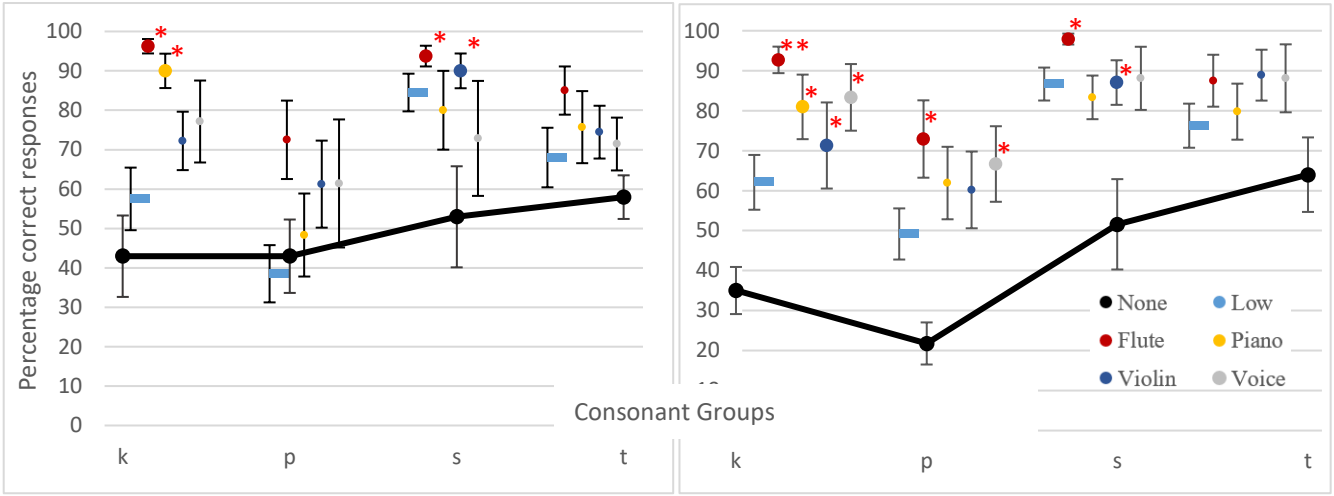
In applying a post-hoc test to the interaction between Part and Instrument, instrumentalists show significant differences with participants with no musical experience in part 1 and part 3. In part 1, Flute > None (p <.001), Violin > None (p =.047) and Flute > Low (p <.001). In part 3, every instrument group performs significantly better than participants with no musical experience (None), ps <.01. We also observe that flutists perform significantly better than participants with a low-level of musical experience (Flute > Low, p =.003). The only instrumentalists that show a learning effect (P3 > P1) are singers (64.29% to 97.86%, p =.031). In the interaction Instrument*Consonant, we find that the various musical instrument specializations show different consonant advantages. Flutists shows significant differences with participants with no musical experience (None), where Flute > None (ps =.001) for /k/, /p/ and /s/, and with participants with a low-level of musical experience where Flute > Low (p <.001) for /k/ and Flute > Low (p =.052) for /p/. This gives the following consonant hierarchy: /s = k (=t)> p (=t)/ (p <.001). For violinists, we find an advantage over non-musicians for /k/, Violin > None (p =.032), and /s/, Violin > None (p <.001), showing that /s = t > p/ (p<.001) and /s > k/ (p =.004). In addition, we find that singers and pianists only show advantages for /k/ where Piano > None and Voice > None (ps <.001), though the two instrument groups show contrasting consonant hierarchies, where /s = t = k > p/ for pianists, and /t = s = k = p/ for singers. Finally, we find a significant difference between two instruments for /k/, where Flute > Violin (p =.026), notably reflected in the consonant hierarchies. These results underline a

difference according to instrument which depends on the consonant, with a particular advantage for /k/ present in every instrument group.

In further detailing the results observed, we considered the double interaction Instrument/Level*Consonant*Part (p =.05). Specific comparisons allow us to differentiate the advantages shown for /k, p/ and /s/ according to instrument and part, where we observe that significant differences are present only between instrumentalists and participants with no musical experience. For /k/, we observe in part 1, that pianists (90%) and flutists (96.25%) are significantly different from non-musicians (43%; both ps <.001), and in part 3, in addition to pianists (80.95%) and flutists (92.7%), we also observed that singers (83.33%) are significantly different from non-musicians (35%: ps <.05). In both parts flutists perform significantly better than participants with a low level of musical experience do (57.5% in part 1 and 62.08% in part 3; ps <.02). However, this is not the case for any other instrument, nor any other consonant. For /p/, we observe significant differences only in part 3 for flutists (72.92%) and singers (66.67%) compared to non-musicians (21.67%, both ps <.05). For /s/, we observe significant advantages for flutists and violinists in part 1 (93.75% and 90% respectively) and in part 3 (97.92% and 87.04% respectively) compared to participants with no musical experience (53% in part 1 and 51.67% in part 3; ps <.02). /s/ is also the only consonant for which we also observe a significant difference for participants with a low-level of musical experience and non-musicians (Low > None; p =.032). Although there is no consonant-specific learning effect, we observe that the instrumental advantages are more numerous in part 3 than in part 1, see Figure 3.

Figure 3:

Differences between consonants per instrument and per part (Left- Part 1, Right- Part 3). Significant differences with None are indicated with one star, those with both None and Low are indicated with two stars.



Discussion

Musical advantage as a transfer process

In this study, we took an interest in the role of musical expertise on the perception of whistled speech, while seeking to understand the transfer of skills between music and speech. Firstly, it appears that participants categorize all the consonants well over chance, at 69% (with chance at 25%). When we explored the differences in results according to musical experience, we observed that even a small amount of musical experience provided an advantage for participants. Indeed, participants with only a low-level of musical experience (levels 1, 2 and 3) have a 19.7% advantage over participants with no musical experience. Participants with high levels of musical experience (levels 4, 5 and 6) also showed a 31.6% difference with participants who had no musical experience, and an 11.9% difference with participants who had a low level of musical experience. These results suggest that, in the case of the whistled speech mode, transfers between music and speech occur with even a small amount of musical

knowledge. However, these differences are stronger and more apparent with a higher level of musical skill. As these levels were divided according to skill rather than years of musical training (or other criteria), we suggest that though the number of years of musical experience may affect the musician participants' results (for example in Gordon & Magne, 2017), the musical skill level achieved by participants is a more interesting measurement. Indeed, it allows us to target possible musical knowledge and its effect on speech perception.

The differences between the three musical experience groups were underlined in the analysis of parts and consonants. In both parts 1 and 3, high-level musicians showed a significant advantage compared to participants with no musical experience (24.3% difference in part 1, and 37.6% in part 3), as well as with participants who had a low-level of musical experience (a difference of 12.2%). In addition, high-level musicians demonstrated a learning effect of 14.53%. This seems to suggest that musical training affects the ability to learn and adapt to whistled consonants, even in a context containing more variability (in part 3). We could argue that this ability to learn is also visible for participants with a low-level of musical skill, as they show significant differences with non-musicians in part 3 but not in part 1.

These differences are further specified in the analysis of whistled consonants. Although all participants categorized whistled consonants well over chance, high-level musician participants showed an advantage for 3 of the 4 consonants compared to participants with no musical experience: /k/, with 44.67% difference, /p/, with 28.69% difference and /s/ with 33.79% difference. Participants with a low-level of musical experience also showed an advantage for /s/ of 24.89% compared to participants with no musical experience. Interestingly, by separating participants into three groups, the advantage for /s/ noted in previous experiments (Tran Ngoc et al., 2020a; 2022a) only appears once participants have a

bit of musical experience. This suggests that the combination of pitch change and stop is clearest for participants with no musical experience, whereas the distinct perception of pitch change combined with a dip without clear stop (like /s/) improves with musical training. This is surprising as we noted that /s/ reaches slightly higher frequencies than /t/, thus amplifying the pitch change. However, it could be that the emulation of the fricative noise in the "dip" creates a more delicate shift in the sound quality of the signal (and thus its perceived 'timbre'), better recognized with musical experience. The advantage shown by high-level musicians is strongest for /k/, suggests that these participants may use the "sharp amplitude modulation" cue in the consonant categorization much more than other participant groups. This suggestion is supported by the learning effect found for /t/ (specific to high-level musicians) which also uses a "sharp amplitude rise". The advantage shown for this specific acoustic cue could also explain the performance given for /p/, which opposes /k/ only in terms of articulatory cues reflected in amplitude dynamics. Another explanation for this could be a better awareness of articulatory cues compared to participants with little or no musical experience. These differences show that consonant cues are recognized and exploited differently according to musical experience, where only a high-level of musical experience helps participants to focus on acoustic details akin to timbre and the articulatory related cues.

Instrument specificity and identifying transfer mode

Though these findings underline a transfer from musical knowledge to whistled speech perception, they reveal little about the capacities that led to these advantages. When including instrument specialization in our analyses and opposing 4 different instruments from different families, each with their own skill set, we further detailed the transferred skills according to the differences observed between each instrument group. Flutist participants

show the strongest advantage compared to participants with no musical experience. This is not only the case for overall performance rate (41.4% difference), but also for every consonant: in each part for /k/ (53.25% difference in part 1 and a 57.7% difference in part 3) and /s/ (40.75% difference in part 1 and 46.25% difference in part 3), and in part 3 for /p/ (51.26% difference). Flutists also show a significant advantage over participants with a low-level of musical experience for /k/ (in both parts), and for /p/ more generally, thus specifying the general high-level musician advantages described previously. We can explain the flutists' advantages in several ways. We first suggest that the similarity in sound quality between whistled speech and the flute may help flutists to identify the essential acoustic cues. This capacity for enhanced sound categorization according to timbre reflects previously demonstrated timbre-based advantages shown in other contexts (notably for cortical representations of tone, or other neural activity, see for example Margulis et al., 2009; Shahin et al, 2008). Second, the similarities in production, notably the use of consonant articulation in flute attacks (see Dickey & Lasocki 2020), could give flutist participants a more expansive awareness of these consonant productions, and the possible variations present. This may explain the differences found for /k/ and /p/, however, it would be interesting to explore whether this perceptual capacity is present for other wind instruments whose timbres are not so similar to that of whistled speech.

Other significant differences are observed between instrumentalists and participants with no musical experience, thus differentiating each instrument. Pianists showed advantages for /k/ in part 1 (of 47%) and in part 3 (of 45.95%), singers showed an advantage for /k/ and /p/ in part 3 (of 48.33% and 44.99%), and violinists showed advantages for /s/ in parts 1 (of 37%) and 3 (of 35.37%). Singers also demonstrate an overall learning effect (18.24% difference between parts), suggesting that each instrument has its own categorization profile and

behavior. This suggests that though flutists may show advantages due to both timbre and articulation, the contribution of these two traits are quite diverse according to musical expertise. It is therefore difficult to generalize such musical advantages as homogeneous when different skill-sets show specific advantages for certain consonants.

The differences between instrumentalists' results show that the transfer between music and speech varies according to the instrument profile, suggesting that specific changes in auditory perception take place according to the instrument played, rather than through domain general modifications due to musical training. This supports the waterfall hypothesis, implying that, for musician participants, the instrument-specific, low-level perceptual-cognitive changes allow for improved performances rather than commonly trained musically skills such as memory, attention or executive functions. Indeed, though these skills may also be more efficient for musicians, we suggest that the advantages found in this consonant categorization task do not stem from them. Finally, as these results are based on whistled speech, which only amplifies certain aspects of the speech signal, they are difficult to reproduce with modal speech. Nonetheless, we suggest that a more generalized application of this experiment towards other forms of modified speech, with different categories of acoustic cues, would help us understand the effect of musical experience more generally.

Conclusions

This study on whistled consonant categorization allows us to take an interest in the transfer process between music and speech. In addition to determining the skill level needed to show a musical advantage in whistled speech perception, we also tested two different transfer hypotheses, one based on auditory perception and the other on cognitive skills. The results obtained confirm the advantages of musical experience, even at a low musical skill

level, but show that higher-level musicians have a stronger advantage, notably in the final part of the experiment and for certain consonants. When we further nuanced this qualification through instrument specialization, we underlined a difference in results according to instrument. We suggest that these results confirm the waterfall hypothesis, where the transfer between speech and music would first apply to the perceptual level, based on the instrument skills learned (listening and production) rather than improved memory or executive functions.

Chapter 5

Musical experience and the whistled word

Introduction

In this final chapter, we take an interest in the effect of musical experience on the whistled word through a single article and behavioral experiment (Expt 8). Having established how participants with musical experience show a large advantage over participants without musical experience in the phoneme, this final experiment considers whistled word recognition and phoneme correspondence according to musical expertise.

The behavioral experiment presented here reprises the 1-part structure of the whistled word experiment presented in Chapter 3 (Expt 5), using the productions of a single whistler (with intra-whistler variability). In doing so, we approach the complexity of the word-based stimulus by choosing to focus on the phoneme among other various possible pre-lexical units. This choice of target unit was essential in constructing the whistled stimuli, thus excluding other prelexical units. We also choose to approach musical experience in this article by reprising the groups proposed in Chapter 4.2, reflecting an evolution in thinking around our definition of musical experience throughout this thesis, which could be reconsidered in further studies.

When considering musical experience in this chapter, already largely discussed in Chpt.1.3, and more briefly in the context of Chapter 4, we can compare the proposed

groupings and levels used in Chpt.4.1 and Chpt.4.2 with the other possible criteria. Indeed, our initial definition of the musician, as used in Chpt.4.1, is someone who has mastered technical skills, but who also associates with such a musical identity (Mills, 2010). This identity, both in terms of skill (Gracyk, 2003) and through musicianship (Hargeaves et al., 2011) underlines a socially defined concept of the musician and the skills involved. This binary opposition is coherent with most experimental work analyzing musical experience (see Annex A.1.). However another, perhaps more inclusive manner of understanding and defining musical experience is by considering a continuous increase in musical skill. This was used to create the three groups in Chpt.4.2 (Non-musician, Low-level Musician, and High-Level musician). As underlined by Smit et al. (2023), the criteria for defining a musician lacks homogeneity (with criteria sometimes including years of experience, starting age, musical exams, or standardized tests, among others) and they suggest that musical experience and knowledge may be more continuous rather than threshold-based. As this was true in the whistled consonant categorization task (Expt 7), we chose to reprise the musical levels proposed in that experiment, based on the French conservatory system. In doing so, we can also consider musical instrument specialization for high-level musicians, including the same target instruments as used in Chapter 4: violin, piano, flute, and voice (see Chapter 4, Introduction). As these musical skills are shown to impact whistled word perception, with a larger advantage shown for high-level musicians, we also included participants with expert knowledge of whistling in a secondary analysis proposed in this article. We compared performances between these two groups of experts by taking an interest in phoneme correspondence, which allowed us to compare the results from previous experiments according to specific phonemes with those of the word (for both musician participants and expert whistlers).

In sum, in this experiment, we chose to target the whistled phoneme in the context of the word and to test musical experience according to three groups in order to consider the continuous evolution of skills. This is summarized in Table 7.

Table 7:

Description of the experiment in the article of Chapter 5

Chapter 5 - Words	**Articles**	**Expt**	**Target**	**Ref.**	**Participants**	**Groups** 0	1 2 3	4 5 6	
5.1	Tran Ngoc et al., 2023d	Expt 8	Musical experience and whistled words	Expt 8	67+7	None	Low	High Inst	Expert Whistlers

5.1 Musical experience and speech processing: the case of whistled words

Abstract

In this paper, we explore the effect of musical expertise on whistled word perception by naive listeners. In whistled words of non-tonal languages, vowels are transposed to relatively stable pitches, while consonants are translated into pitch movements or interruptions. Previous behavioral studies have demonstrated that naive listeners can categorize isolated consonants, vowels, and words well over chance. Here we take an interest in the effect of musical experience on words while focusing on specific phonemes within the context of the word, considering the role of phoneme position and type, and comparing the contribution of these whistled consonants and vowels in word recognition. Musical experience shows a significant and continuous advantage according to the musical level achieved, which, when further specified according to vowels and consonants, shows stronger advantages for vowels over consonants, and high-level musicians over non-musicians. By specifying high-level musician skill according to one's musical instrument expertise (piano, violin, flute, or singing), and comparing these instrument groups to expert whistlers, we observe instrument-specific profiles in the answer patterns. The differentiation of such profiles underlines a resounding advantage for expert whistlers, as well as the role of instrument specificity when considering skills transferred from music to speech. These profiles also highlight differences in phoneme correspondence rates due to the context of the word, especially impacting "acute" consonants (/s/ and /t/), and highlighting the robustness of /i/ and /o/.

Article Information

Article Status

This article is currently under review for Cognitive Science.

Tran Ngoc, A., Meyer, J. & Meunier, F. (2023d). Musical experience and speech processing: the case of whistled words. *Cognitive Science*.

Keywords: speech perception, knowledge transfer, whistled speech, musical experience

Data link: https://osf.io/x5c4k/?view_only=294acfef0ff84dfe92caf0bb4c979ed2

Musical experience and speech processing: the case of whistled words

Introduction

Musical experience affects speech processing in various ways and on various levels (see review by Besson et al., 2011). This includes better performance in processing at phonological levels, notably in foreign speech, both in phonological production (Milovanov et al., 2010) and phonological perception (Slevc & Miyake, 2006). Such advantages have also been shown to extend to modified speech conditions such as speech in noise (Varnet et al., 2015; Bidelman & Krishnan, 2010), due to a better representation of the target acoustic stimuli (Parbery-Clark et al., 2009) or improved attention skills (Strait & Kraus, 2011). Such findings underline a similarity in processing between speech and music (see review by Sammler & Elmer, 2020), as well as common structural aspects.

A promising new path combining language processes and musical expertise considers musical surrogate languages to understand shared processing mechanisms (McPherson & Winter, 2022). Along the same line, recent studies testing language processing by musicians have been applied to whistled speech (Tran Ngoc et al., 2023a; Tran Ngoc et al., 2022b). This practice, used to transpose acoustically spoken dialogs rather than as a type of musical production, reduces the vocal spoken signal to a simple modulated whistled line akin to a musical melody. Whistled speech has evolved in a large diversity of languages worldwide in mountainous and densely forested regions, enabling true distance communication. The physical characteristics of whistles are well adapted to the acoustic limitations in the environment, as they focus on a narrow range of frequencies (1,000–4,000 Hz) that favor sound propagation and that are higher than most prevalent natural background noises (emphasizing low-frequency contents). Moreover, these frequencies are optimal for human

audibility and sound discrimination (Meyer, 2021a). The transposition of the linguistic segments – typically vowels and consonants – by speakers of non-tonal languages (such as Spanish, Turkish, Tamazight, and Greek) is one of the most interesting aspects of whistled speech, proposing alternative insights into how the acoustic realization of phonemes can be drastically reduced without hindering recognition from listeners. While mastering this speech form does require training, recent studies have shown that even without extensive training, naive listeners can successfully categorize phonemes in both whistled consonants and whistled vowels. Recent findings demonstrate how naive listeners can already categorize phonemes in this modified form correctly and well over chance (Tran Ngoc et al., 2020b; Tran Ngoc et al., 2022a). These categorization tasks highlighted differences in performance depending on the consonants and vowels heard (among those of interest).

In such languages, whistled speech produces different pitch categories according to the spoken vowel timbre, thus transposing each of the vowels of modal speech to a specific whistled frequency range (which is also relative to the speaker and the whistling technique). In whistled Spanish, the language tested in these previous studies, the whistled vowel pitches can be ordered from highest to lowest in the following manner: /i/, /e/, /a/, and /o/, with /u/ generally overlapping with /o/ and /a/ (Busnel & Classe, 1976; Rialland, 2005; Meyer, 2015). Whistled consonants modulate/change these pitches according to their corresponding spoken articulation. For example, in the VCV context with /a/, articulation can cause a large and rapid pitch change for consonants /s/ and /t/, or only a minor pitch change for /k/ and /p/. We also observe an opposition between continuous or near-continuous consonants (like /s/, considered semi-continuous, or /p/) and clearly interrupted consonants (/t/, k/). This reflects various cues present in modal speech (see Meyer, 2015; Diaz, 2008; Tran Ngoc et al., 2022a). These acoustic cues have been used to characterize, categorize, and regroup whistled

consonants, using the opposition between “continuous” and “interrupted” consonants, and “acute” vs. “grave” consonants (Trujillo, 1978; Rialland, 2005; Diaz, 2008). The latter distinction is based on acoustic loci (high vs. low) that mimic those of the spoken word. These lead to whistled transitions with distinct characteristic patterns that resemble spoken formant transitions. These patterns are influenced by the surrounding vowels (Leroy, 1970, Rialland, 2005; Meyer, 2015).

The ability for phoneme categorization in whistled speech by naive listeners has also been extended to words: Tran Ngoc et al. (2023a) showed that whistled words could be recognized well over chance (which was at 20%), with 45.6% of correct responses obtained. However, when compared to phoneme categorization rates, participants did not show significant improvements in word recognition. This contrasts with the performances of expert whistlers, where previous experiments have shown word categorization to be closer to 60-75% in whistled Turkish (Busnel, 1970), with an increase of 20-30% of correct answers compared to VCV or CV tokens (Meyer, 2015). Rather, Tran Ngoc et al. (2023a) highlights differences for consonant and vowel recognition rates, where vowels were much better recognized. The vowel hierarchies deduced from these results were generally consistent with the hierarchies found with isolated vowels.

In line with these results, we propose using whistled speech as a tool to understand speech perception because this type of speech induces a different perception of fully intelligible words or sentences. This change has sometimes been interpreted as an example of “perceptual insight” or of a pop-out in a top-down perceptual process (Meyer et al., 2017), where higher-level knowledge and expectations apply to sounds that can potentially be heard as speech [much like what happens in artificial Sine Wave Speech (see Remez et al., 1981;

Davis and Johnsrude, 2007)]. Here, we choose to study the whistled word, thus reprising previous considerations concerning the role of phonemes in words (Benki, 2003; Delle Luche et al., 2014; Tran Ngoc et al., 2023a) and whistled word perception as a top-down/bottom-up process. We wonder how musical experience will affect the relationship between phonemes and words.

Indeed, the benefit of musical training on speech perception is sometimes considered to be general, as, like in speech, musical experience engages cognitive processes like attention, memory, and executive functions (Schellenberg & Weiss, 2013; Parbery-Clark et al., 2009), potentially leading to improved performances in other tasks. However, more specific transfers have also been considered in the literature, as musical training also involves learning to identify and distinguish elements in auditory stimuli such as different pitches, rhythms, and tones, which are also present in speech. Such skills could therefore be transferred to the perception and processing of speech sounds. Peretz and Coltheart (2003) propose a mechanistic model for sound processing that supports this form of transfer, where sounds are initially treated according to a “common acoustic analysis" before feeding into either a music-specific module ("contour analysis"), a language-specific module ("acoustic-to-phonological conversion"), or an as-yet-uncharacterized module ("rhythm and meter"). The shared perceptual capacities and the initially common acoustic analysis could explain a potential crossover between the two perceptual systems, leading to certain advantages. In addition, as each musical instrument has specific acoustic properties, such transfers could vary depending on the instrument played.

The unique form of whistled speech also allows us to consider the role of articulatory and acoustic cues in speech processing, a crucial issue that opposed Motor with Acoustic theories of speech perception. In Motor Theory, see Liberman et al. (1967) and Liberman &

Mattingly (1985), speech perception is based on the matching of articulatory gestures to one's own articulatory representation of sound, thus relying on a knowledge of speech production for perception. However, according to Acoustic theories, speech perception would use acoustic cues as tools for speech perception without considering production (Fant, 1960). In whistled speech, though the articulatory cues found in modal speech are used in production, the acoustic realization of these forms reduces the complex signal of modal speech. This therefore modifies the relationship between articulatory and acoustic cues found in modal speech.

Here we target the effect of musical experience and different types of instrumental specialization more specifically, comparing naïve listeners to participants with a knowledge of whistled speech. These forms of experience add complexity to the participants' relationships with acoustic and articulatory cues. This leads to several possible analyses, enabling us to explore the relationships between cues with different focuses/insights. Indeed, participants who have experience with whistled speech have a full knowledge of whistled articulatory and acoustic cues, which is similar to their knowledge of modal speech, even though they may be unfamiliar with the whistled words heard. This contrasts instrumental knowledge, where musical skills have been shown to transfer to other auditory skills. Moreover, on musical instruments, the production of the sound is very different from speech production. Wind instruments are a slight exception to this, as production of attacks often use spoken articulation. These forms of experience therefore enhance the divide between the different roles played by acoustic and articulatory elements in whistled speech perception, providing an insight into these theoretical speech perception models.

In this study, we focus on the same whistled French words as Tran Ngoc et al. (2023a) and consider the effect of musical experience on word recognition and phoneme correspondence. In doing so, we assume that the natural, yet modified, whistled speech form represents a relevant tool to investigate perceptual processes in language, and more specifically the impact of expertise – such as musical experience – on speech perception. We then measured speech processing according to musical instrument specialization (for violin, piano, flute, and voice) to detect differences in perception according to specific instrumental production and perception skills, and compared this form of expertise to that of expert whistlers.

We ask the following questions: What is the effect of musical experience on whistled word perception? What is the contribution of specific vowels and consonants to word recognition? If an advantage exists between participants with musical experience compared to non-musicians, how do expert musicians' skills compare to those of highly trained whistlers? Are musical advantages for whistled words specific to individual instrument specializations? Finally, how do these results reflect certain theoretical approaches to speech perception?

In this paper, we address these questions through the presentation of a single behavioral experiment studying whistled word categorization. We include two different analyses: one which considers participants according to three groups of musical skill level (None, Low-level, and High-level), and another which compares only the high-level musicians according to targeted instrument specialization, and a group of expert whistlers, all teachers in whistled speech.

Experiment

Method

Stimuli

Whistled Words

We selected 24 French words for this categorization task, chosen to include vowels and consonants from previous experiments, thus enabling us to compare results.

The words selected were disyllabic nouns with a CVCV(C) structure, noted as C1 V1 C2 V2 (C3). These words included only the vowels of interest [i], [e], [a] and [o], which were equally represented in each vowel position, each appearing 6 times as the V1 and 6 times as the V2, providing two occurrences of each V1-V2 combination (a-o, a-e, a-i, o-a, o-e, o-i, e-a, e-o, e-i, and i-a, i-o, i-e). We also selected words that included the four consonants used in previous experiments, [k], [p], [s], and [t]: each appeared both at the start of the word (C1 position), for at least 4 words, and in the second consonant position, for 3 words (C2 position). To ensure that words were known by all participants, we controlled their frequency of apparition in an adult lexicon (Lexique by New & Pallier, 2023). The frequency of occurrence out of 1 million words averages 55.31 (SD = 180.25). The completed word list (see Annex, A.5.2) fulfills these criteria, though to do so, several other consonants were also present ([b, d, f, ʃ, m] in the initial C1 position and [ʃ, n, l, m, g, ʁ, d, z] in the C2 position).

In adhering to this criteria, and as each word was recorded 4 times, the target consonants /k/, /s/ and /t/ appear 16 times and /p/ appears 20 times in C1, and each of the target consonants (/k/, /p/, /t/ and /s/) appear 12 times in C2. A single whistler, fluent in whistled Spanish and sufficiently knowledgeable in French to properly pronounce the words,

was recorded on a Zoom H1. It should also be noted that the target vowels and consonants included in this test have similar pronunciations in both French and Spanish. The recordings nonetheless consisted of a spoken version of the word (used to control the pronunciation) followed by the whistled version which was repeated 4 times.

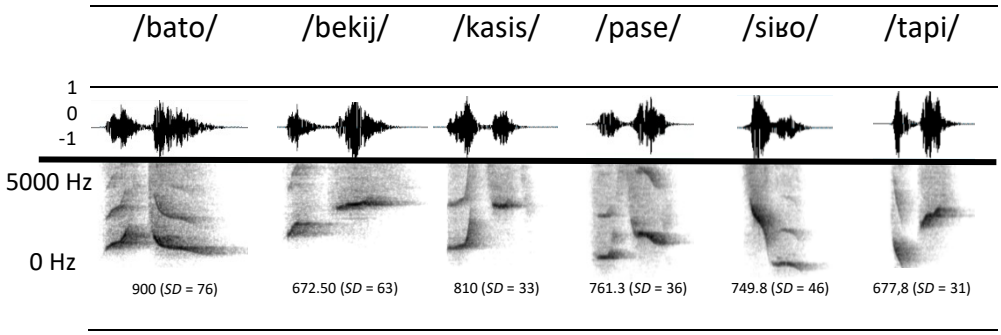
In the transformation from usual modal spoken speech to whistled speech, the salient characteristics of the spoken word are reflected in the whistled pitches and amplitude modulations. Indeed as previously observed (Tran Ngoc et al., 2023a), the duration of the word in whistled speech (i.e. its elongation) is not correlated with the word duration in modal speech, though it is in agreement with French prosody. While each whistled vowel is produced within a certain pitch range, in the context of the word, the position of the vowel affects the variability of the pitch range (or the stability of the vowel) – where the V2 vowels /e/, /a/ and /o/ are more stable than the corresponding V1 vowels. It also appears that the V1 /o/ is much higher than the V2 /o/ (see Tran Ngoc et al., 2023a for a more detailed description). For consonants, it appears that the consonant cues described in the VCV format (see Rialland, 2005; Trujillo, 2006; Tran Ngoc et al., 2022a) are drastically modified in the context of words, because of elements of co-articulation.

In the C1 position, distinctions between “continuous” consonants and “interrupted” consonants become superfluous, lacking the preceding vowel. Thus, the C1 consonant is better characterized by pitch change (see for example /siʁo/), though we could also consider amplitude rise (for example in /tapi/) or articulation points (bilabial - /p/ or glottal - /k/, and dental - /t/). In the C2 position, descriptions remain more consistent with previous studies, once again including the opposition between continuous (or semi-continuous /s/) and interrupted (/k/, /t/, /p/). However, the “acute”/”grave” opposition is also affected by the

vowel context, modifying the size of the pitch change of acute consonants (for example in /kasis/ and /pase/, see Table 1).

Table 1:

Wavelength and spectrogram of whistled words representing /k, p, s, and t/ in C1 and in C2



Design

The design for this experiment is identical to that of Tran Ngoc et al., 2023a. This experiment included a main portion, where we evaluated participants' recognition performance for the 24 whistled words. Each word was produced four different times and therefore included natural production variation. For each word heard, participants were proposed five word options. Given the novelty of this word categorization task, we chose to propose a limited amount of word options to the participants, also maintaining continuity with the previous vowel-focused study. These options included four-filler words, which were selected randomly (using https://www.random/org/lists/). We constructed two answer lists, which were randomly attributed to each participant. The experiment was conducted online and was programmed using PCIbex Farm. Thus, participation took place at home using headphones, earbuds, or speakers.

Procedure

Before starting the experiment, participants answered a short questionnaire indicating their native language, their age, and gender. A detailed description of their musical experience

was also requested, including the instrument played, a self-evaluated musical "level" and their background in the said instrument (number of years played/context). We asked participants to choose between six different musical levels: 1 – beginner (*Débutant*), 2 – amateur, 3 – confirmed (*Confirmé*), 4 – DEM musical Diploma (*DEM*), 5 – Superior University Diploma (*Diplôme Supérieur*), and 6 – professional musician (*Professionnel*). If participants had no musical experience, they were asked to leave this section blank.

Once the questionnaire was completed, the experiment format was presented to participants by showing an example of a whistled word, /pate/ ("pâté"), as well as the drop-down answer menu. When the experiment began, participants heard a randomly selected whistled word from the list and had to pick the corresponding word among the five choices suggested (which included the correct answer) from the drop-down menu. They were then asked to validate their answer, and the next whistled word was played immediately afterward. Thus, participants first heard the word and then viewed the possible responses. Each of the 24 words was presented four times (using four different recordings), for a total of 96 words heard.

Participants

Ninety-three participants were included. They were all French speakers and had no language impairments or hearing problems. The participant group included 53 women and 40 men, who were between 18 and 50 years old, with an average age of 27.52 years old (SD= 6.18). Within this group, 18 participants had no musical experience whatsoever, 2 participants declared having a "beginner" level, 11 participants declared being "amateur" musicians, 22 participants declared being "confirmed" musicians, 16 participants had obtained the "DEM"

or “Musical Studies Diploma”, 6 participants had obtained the “Superior Musical Diploma”, and 18 participants were professional musicians.

In addition, and in order to have a control group, we asked 7 expert whistlers with low levels of musical experience to complete the task. Four participants had no musical experience, one was a beginner and two were amateur musicians. This whistler group consisted of native Spanish speakers with a basic knowledge of French, who had no language or hearing problems.

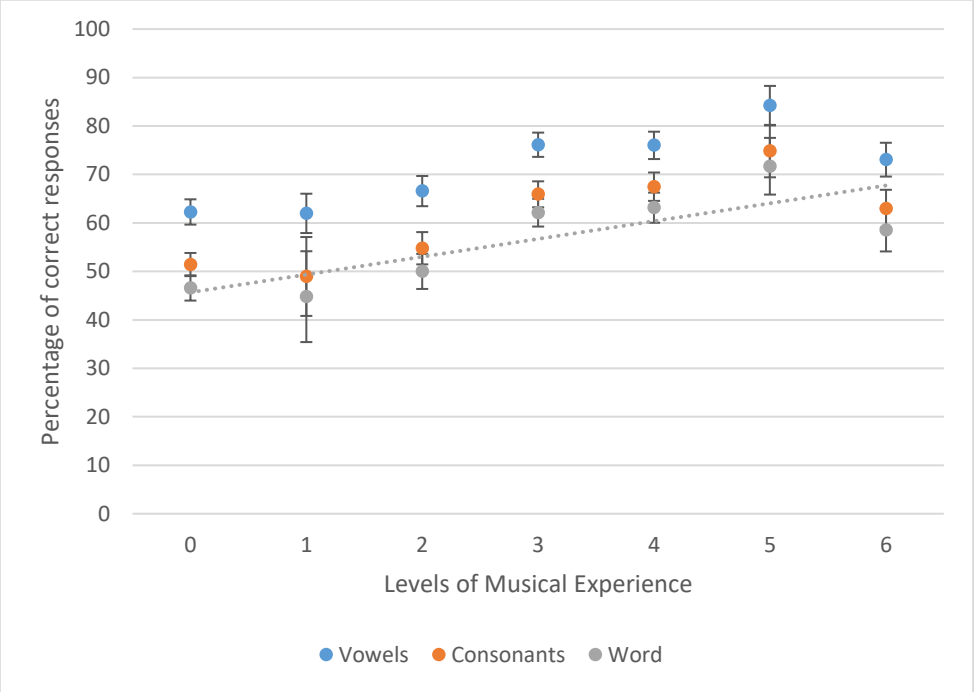
This experiment was conducted in accordance with the Helsinki agreement.

Results

In a first analysis, we included only the 93 naive French speakers. Overall, participants categorized whistled words correctly with an average of 57.4% of correct responses obtained (*SD* = 15.53), well over chance at 20%. When considering the percentage of correct responses obtained per word according to musical experience (Levels 0, 1, 2, 3, 4, 5, 6), we observe a strong increase in overall correct answers according to Level, going from 46.58% at Level 0 (*SD* = 10.68), to a maximum of 71.70% at Level 5 (*SD* = 14.33). This progressive increase shows a significant positive correlation between level and correct answer rate per word (Pearson’s correlation r(91)= 0.36, *p* <.001), suggesting that there is an advantage for participants with musical experience, see Figure 1. We can further investigate the effect of musical experience at the phoneme level.

Figure 1:

Percentage of correct word responses and phoneme correspondences (all phonemes) with standard error, obtained according to levels of musical experience, with trend for word responses



To consider the effect of musical experience, we differentiate participants with no musical experience (Level 0), called “None”, from participants with some musical experience (Levels 1, 2, and 3), called “Low”, and participants with a high level of musical experience. We define this final group (called “High”) through the completion of a musical diploma (Levels 4, 5 and 6). We observe that even the group None obtained correct word answers above chance, with 46.58% of correct answers (*SD* = 10.9). Participants with a low level of musical experience (Low) recognized words with 58.08% of correct answers (*SD* = 14.04), and participants with a high level of musical experience (High) obtained 62.37% (*SD* = 15.5). The amount of correct answers obtained therefore increases according to musical experience.

We then considered both correct and incorrect answers, and the correspondence between phonemes in played and answered words according to position (1 or 2). We focused on the phonemes of interest (/a,e,o,i, k,p,s,t/). There were 96 vowels played in V1, and 96

vowels played in V2 for each of the 93 participants. Therefore, we considered 17856 elements of vowel data points. For the consonants, those of interest (/k,p,s,t/) appear a total of 68 times in C1 and 48 times in C2 for each of the 93 participants, amounting to consonant 10788 data points. Thus in total, we considered 28644 data points. We applied a Generalized Linear Mixed Model (GLMM) to phoneme correspondence, with Phoneme Type (Consonant, Vowel), Position (1, 2), and Musical Experience (None, Low, High) as fixed factors. We included Participants and Words as random effects.

We find a significant main effect of Phoneme Type (X^2(1, N = 93) = 41.019, *p* <.001) showing that correspondence rate is higher for vowels (at 68.4%) than for consonants (at 64.1%). We also observe a significant effect of Musical Experience X^2(2, N = 93) = 15.095, *p* <.001) and a significant interaction between Phoneme type*Musical Experience (X^2(2, N = 93), *p* =.039). Post-hoc tests (using Bonferonni correction) reveal that the difference between vowel correspondence rates (V) and consonant correspondence rates (C) is significant only for musicians: low-level musicians (V = 69.1% vs. C = 63.1%) and high-level musicians (V = 72.3% vs. C = 68.8%; *p*s <.001). This is not the case for the group of non-musicians. We observe significant differences between High and None for both vowels (72.3% vs. 58.3%; *p* <.001) and consonants (68.8% vs. 55.6%; *p* =.006). For vowels, we also observe a tendency for difference between Low and None (69.1% vs. 58.3%; *p* =.054). Overall we did not find an effect of Position (*p*s >.05).

These results suggest that musical advantages for vowel and consonant recognition within the word increase with experience, with a stronger advantage for high-level musicians (defined through a musical diploma) over non-musicians, than for low-level musicians (defined through self-evaluation) over non-musicians. In light of these results, we wish to further

explore the effect of a high-level of musical experience by defining high-level musicians' knowledge according to their instrument specialization.

In a second analysis, we focused only on participants with a "high-level" of musical experience, to explore the impact of instrument specialization more precisely. We targeted four instrument groups among the high-level musicians: violin, piano, flute and voice; and retained only the high-level musician participants who played these instruments, reducing the number of participants from the previous analysis. Among these high-level musicians, 6 were singers (Voice), 7 were flutists (Flute), 8 were pianists (Piano) and 7 were violinists (Violin). In addition, and in order to better characterize the performance of instrumentalists, we included the performances of a group of 7 expert whistlers with low levels of musical experience who have a fluent knowledge of whistled speech in Spanish (Silbo). This amounts to a total of 35 participants.

In this second analysis, we also considered the eight phonemes included in the words: /a,e,i,o,k,p,s,t/. We did not include the factor Position as this showed no significant effect in our first analysis. We applied a GLMM to phoneme correspondence with Phoneme (/a,e,i,o,k,p,s,t/) and Group (Violin, Piano, Voice, Flute, Whistler) as fixed factors. We included Participants and Words as random effects.

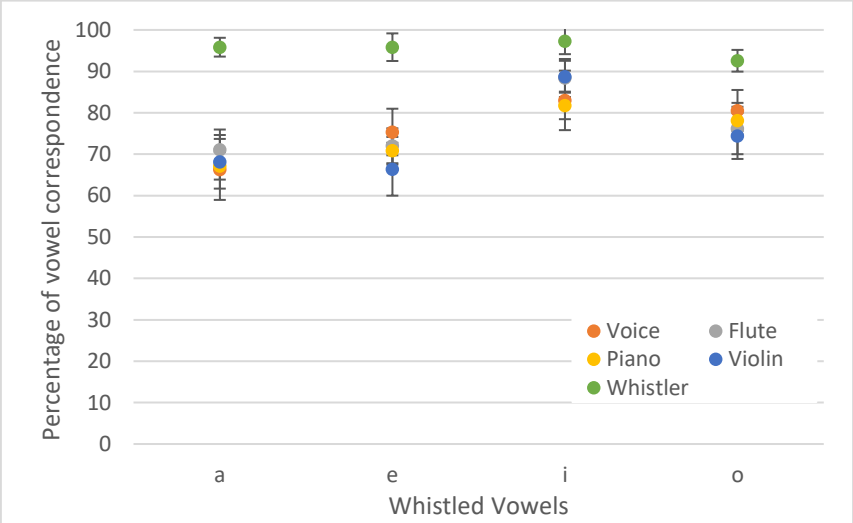
We find a significant main effect of Phoneme ($X^2(7, N = 35) = 79.6$, $p < .001$), where the correct correspondence rates of the vowels are at 87.8% for /i/, 80.3%, for /o/, 76% for /e/ and 73.8% for /a/. The consonant correspondence rates were at 74.7% for /k/, 74.1% for /t/, 73.8% for /p/ and 67.4% for /s/. We also observe a significant main effect of Group $X^2(4, N = 35) = 39.4$, $p < .001$). When considering the performance of each of the groups, we observe that overall flutists obtain 75% correct phoneme correspondences, singers 73.4%, violinists

70.9%, and pianists 70.7%. The whistlers show a much higher performance rate with 94.7% correct correspondences obtained. The Phoneme*Group interaction is significant (X^2(28, N = 35) = 66.9, *p* <.001), and we applied post-hoc tests to specific comparisons of this interaction, using the Bonferonni correction.

We observe significantly different profiles for each of the groups present, underlined by differences with the whistlers, see Figures 2 and 3.

Figure 2:

Whistled vowel correspondence for each instrument group



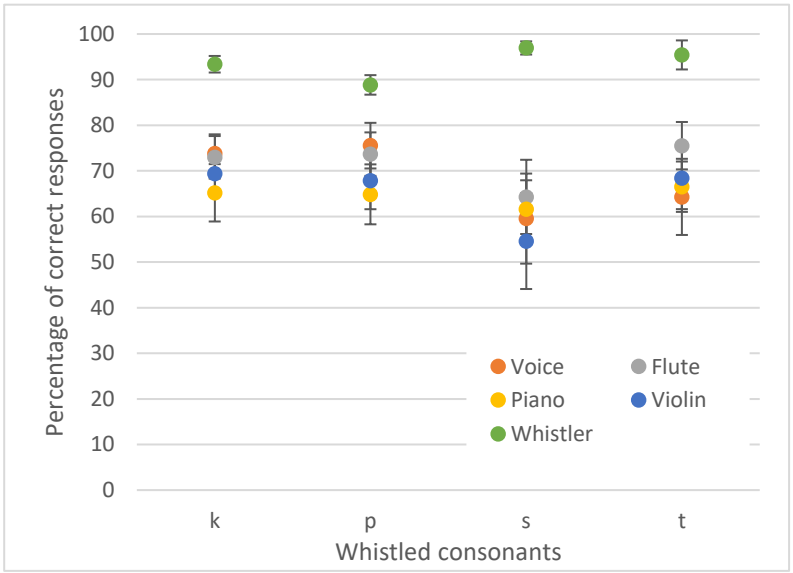
These differences also highlight the presence of some phonological hierarchies according to the group. Compared to pianists, the whistlers perform significantly better for every phoneme except for /o/ (/i/, *p* <.05; /e/, *p* <.001; /a/, *p* <.001; /k/, *p* =.01, /s/, *p* <.001; /t/, *p* =.002). There are no significant differences between phonemes among the pianist group. Compared to singers and violinists, we observe significant advantages for whistlers for two vowels and two consonants: /a/ (*p*s <.001), /e/ (*p*s <.05), /s/ (*p*s <.001) and /t/ (*p*s <.05). There were no significant differences between the phonemes for singers, while, for violinists, we observed significant differences among the vowels, where /i/ is better recognized than each

of the three other vowels: /e/ (p <.001), /a/ (p <.001) and / o/ (p =.004). Finally, for flutists, whistlers show a significant advantage only for two vowels and one consonant: /a/ (p =.031), /e/ (p =.004), and /s/ (p <.001). We also observe significant differences among the vowels within the flutist group, where /i/ is better recognized than /e/ and /a/ (ps <.05) and tends to be better recognized than /o/ (p =.065). Whistlers show no significant differences in performance between or among phoneme correspondence rates.

These differences underline how each of the high-level instrumentalists is significantly different from the whistlers for /a/, /e/, and /s/. Flutists are the only instrument group to show differences solely for these three phonemes. For each of the other instrumentalists, whistlers show significantly more advantages: on /t/ for singers and violinists and on /t/, /k/, and /i/ for pianists, see Figures 2 and 3.

Figure 3:

Whistled consonant correspondence
for each instrument group, shown with standard error



Discussion

In this study, we considered the impact of musical experience on whistled word recognition among naive listeners. Participants perform well over chance (at 20%), with a categorization rate of 57.4%, and we observe an increase in the percentage of correct responses obtained according to level, supported by a significant positive correlation between musical level and correct responses. This led us to contrast three groups of participants according to musical experience while considering phoneme correspondence between played and answered words.

These groups show a gradual improvement in response rates. We find that participants with a low level of musical experience perform better than those with no musical experience (10.74% increase), and those with a high level of musical experience perform better than those with a low level of musical experience by 5%. When comparing these three groups, we also took into consideration the differences between vowels and consonants, extending the analysis applied in Tran Ngoc et al. (2023a), by this time including levels of musical experience. These results, consistent with those obtained previously (i.e. the advantage for vowels over consonants), specify the effect of musical experience according to vowels and consonants. Indeed, low-level musicians show a 10.8% advantage for vowels over non-musicians; and high-level musicians show an advantage over non-musicians of 14% for vowels and 13.2% for consonants. This suggests a continuous improvement in musical skills according to experience and thus a continuous increase in skill transfer according to musical level, as suggested by Smit et al. (2023).

We further explored high-level musicians' behavior by specifying their knowledge according to musical instrument expertise and comparing it with expert whistlers. This

comparison underlined several important differences. Firstly, there was a large performance gap between whistlers and expert musicians, where knowledge of whistled speech (even with a limited amount of experience with French) produced results with almost 100% accuracy (an average of 94.7% of correct phoneme correspondences). Such a performance surpasses any musically related transfer towards whistled word perception, where the highest performing musical instrument group (flutists) obtained a lower score by 19.7%. This suggests that though musical experience may create some perceptive advantages, more targeted training (such as learning to whistle speech and recognize phonemes in whistles) has a much stronger effect.

Secondly, the differences between high-level musicians and whistlers highlighted significantly different profiles according to instrument specialization, characterized by differences in phoneme correspondence rates. Each instrument group showed a different response profile compared to whistlers: flutists showed differences for only 3 phonemes (2 vowels and 1 consonant), violinists and singers for 4 phonemes (2 vowels and 2 consonants), and pianists for 6 phonemes (3 vowels and 3 consonants). These profiles suggest that behavior varied according to instrument specialization, where flutists were most similar to whistlers, and pianists were the least similar. Most notably, in this word recognition task, the flutists behave similarly to whistlers for 3 consonants (/k/, /t/, and /p/). This may reflect instrument-specific similarities in terms of articulation or timbre that do not exist in other instruments. For example, the use of tongued plosives in flute playing with consonants such as /t/ and /k/ resembles that of whistlers, thus highlighting similar production mechanisms. Singers also produce such articulatory consonant movements, however the acoustic sound quality of sung productions are further away from those of whistled speech. Indeed, neither the violin nor the piano emulates speech like sounds while playing, however the musical articulations on the violin produces transitions that are closer to whistled speech, contrary to those of the piano

which are limited due to instrumental constraints, and thus furthest away from whistled speech (Bresin & Battel, 2000).

The difference between phoneme correspondence rates, also shows specificities linked to musical experience. When considering the vowels according to instrument specialization, we underline advantages for /i/. This vowel, systematically categorized better than other vowels in previous studies (Tran Ngoc et al., 2020b; Tran Ngoc et al., 2023a), shows significant differences with the other vowels (/a/, /e/, and /o/) for violinists and flutists (though only a tendency for /o/). Interestingly, difficulties for /o/ in the context of the word (as shown in Tran Ngoc et al., 2023a) also seem present for whistlers, as their performances do not differ from those of musician participants. This could be due to the variability of production of /o/ as documented in the Method section.

When considering the whistled consonant correspondence in the word, all of the high-level musician participants show significant differences with whistlers for /s/, and almost all instrumentalists show a difference for /t/ (except flutists). In addition, all of the instrument groups performed equivalently to whistlers for /p/, and generally behaved similarly for /k/ (except pianists). Interestingly, this consonant hierarchy is reversed compared to the one observed in the VCV context, in which the highest categorization rates are observed for /s/ and /t/ and the lowest categorization rates for /p/ (Tran Ngoc et al., 2022a). Thus, when we consider the whistled cues characterizing these consonants, consonant correspondence in the word underlines a clear opposition between “acute” and “grave” consonants, where “acute” consonant correspondence is more difficult. The difference with VCV consonants may be due to word-specific influences such as vowel co-articulation, producing inconsistent acoustic cues most notably among the acute consonants.

Thus, these findings underline how musical expertise is specific to the type of instrument training received, including knowledge of certain timbres and articulation/production mechanisms. We observe how acoustic improvements through musical experience show an effect on whistled speech perception generally, suggesting that these acoustic cues are important for whistled speech perception. This therefore supports acoustic theories of speech perception. However, the two groups who also use a knowledge of articulatory elements in production of whistled timbres, flutists and whistlers, show greater advantages. This suggests that speech perception also uses articulatory cues in addition to acoustic elements. Thus, instead of favoring the Motor Theory or Auditory/Acoustic theories, we suggest that these findings provide support for theories that establish a connection between articulation and acoustics, such as the PACT theory (Schwartz et al., 2012). Indeed, in such an approach, the perceptive unit is characterized by both its articulatory gestures and its acoustic role. This would explain how improved knowledge of both articulatory and acoustic elements in whistled speech (as is the case for flutists and whistlers) therefore produce the largest advantages.

These results also highlight behavioral differences between phoneme categorization and word categorization, as well as between vowels and consonants. This is first highlighted by the difference between whistled vowel and consonant correspondence rates, where vowel rates are higher than those of consonants. Consonant correspondence rates are, however, much closer to those of the whistled word (see Figure 1). The different treatment of vowels and consonants in the word may suggest that both top-down and bottom-up approaches come into play. Previous findings, such as increased categorization rates for words with larger vowel intervals, as shown in Tran Ngoc et al. (2023a), could suggest a top-down approach to word perception, as identifying the interval considers the relationship between vowels at a

word level rather than as individual phonemes. Higher vowel correspondence rates observed here, which echo findings in modal speech (see Fogerty & Humes, 2010 and Delle Luche et al., 2014), may also rely on the construction of these relationships. However, the similarity between consonant correspondence and word categorization rates could also reflect a bottom-up approach, where the ability to categorize consonants directly affects word choice.

Conclusion

In conclusion, disyllabic whistled word categorization is shown to be impacted by musical experience through a continuous increase in categorization rates, where high-level musician participants show a larger difference in behavior than participants with no musical experience. Interestingly, when comparing high-level musical experience according to instrument specialization with that of expert whistlers, each instrument shows a different profile. These differences generally highlight a very stable response rate for the vowels /i/ and /o/, and a strong impact of the whistled word context on the consonants, where the hierarchy of consonant correspondence is reversed compared to the hierarchy found in the VCV form. Overall, our results clearly show an advantage of musical experience on whistled speech processing for all musicians, with specific profiles depending on the instrument played. This allows us to gain further insight into perceptive approaches used by naive listeners to categorize whistled speech.

Chapter 6

Discussion, conclusions and perspectives

6.1 Overview

In the first part of this discussion, I suggest revisiting the questions outlined in Chapter 1 (1.4). (1) Are naive listeners able to use their knowledge to correctly categorize whistled speech at different perceptual levels (phonemes and words)? (2) What whistled cues are used in this categorization process? (3) Will musical experience have an impact on whistled speech perception? (4) If so, which elements of the speech signal are affected? (5) How do these effects reflect specific skills acquired through musical training such as musical level achieved and instrument specialization?

With these reflections, I hope to highlight how each of these articles and experiments provide a response to these vital questions by giving an overview of the results obtained. In Table 8, we describe the themes incorporated in each chapter and their corresponding experiments. We will use these to retrace the results.

Table 8:

Synthesis of isolated whistled speech experiments

	3-part isolated phoneme experiments					**Words**
Variability	Consonants			Vowels		
General variability	None		Lowered Frequencies	Intra-whistler variability		Intra-whistler variability
Variability in part 3	Inter-whistler variability			Intra-whistler variability (A/B)		N/A
				2 whistlers per expt AAB/BBA	1 & 2 whistlers per expt AAA/BBB/AAB/BBA	
Whistled Speech focus	**Chpt.2.1** Expt 1A	**Chpt 2.2** Expt 1B	**Chpt.2.1** Expt 2	**Chpt.3.1** Expt 3	**Chpt.3.2** Expt 4	**Chpt.3.3** Expt 5
Musical Experience focus	**Chpt.4.2** Expt 7		X	**Chpt.4.1** Expt 6	X	**Chpt.5** Expt 8

6.1.1 Categorization at different perceptual levels

Let us examine the first question: (1) Are naive listeners able to use their knowledge to categorize whistled speech at different perceptual levels (phonemes and words)? We considered this question in every experimental chapter, as both participants with and without musical experience were considered naive listeners - participants who had not previously been exposed to whistled speech. In Chapters 2 and 3, participants had either no musical experience or very little, which contrasted with Chapters 4 and 5 where participants with and without musical experience were included.

Overall, in each of the studies presented, participants categorize whistled phonemes and whistled words well over chance (with chance being defined as 25% for isolated phonemes or 20% in the case of the word). This underlines and confirms the capacity for

correct categorization of whistled speech by all participants. These categorization rates vary according to the conditions of the experiment, such as the whistler, the context, and the listener's musical experience. We will focus more on musical experience (outlined in questions 3, 4 and 5) later on, and can therefore explore question 1 by focusing on non-musician participants (in Chapters 2 and 3). In these two chapters, we considered several perceptual levels: the whistled phoneme (Chapters 2.1, 2.2, 3.1 and 3.2) and the whistled word with a focus on vowels (Chapters 3.3).

Chapters 2.1 and 2.2 focus on whistled consonants and in these chapters, participants obtain an average of 55.1% correct responses (well over chance at 25%) between Expts 1A and 1B. In Expt 1B, participants performed 8.2% better than in 1A. In the lowered whistled consonants of Chpt 2.1 (Expt 2) participants obtained an average of 42% correct responses. The difference in categorization rates between the lowered consonants and non-lowered consonants is also observed when considering the learning effects in these studies. Indeed, we observed a general learning effect only for the lowered whistled consonants (see Chapter 2.1, Expt 2). There was no learning effect in Expt 1A or 1B. Such findings underline the importance of the whistled pitch frequency, where the low-frequency whistled consonants, not as well categorized as the higher-whistled consonants, improve with training to reach rates equivalent to those of the high whistled consonants. This may be because the whistled frequency range (at 1.2 - 3 kHz) echoes the amplified frequencies of the 2nd and 3rd formants.

Chapters 3.1 and 3.2 focus on the whistled vowel in isolation, where participants obtained an average of 53.5% correct responses (well over chance at 25%) in both experiments (Expts 3 & 4). These results indicate that all participants can categorize whistled phonemes well over chance, even for lowered consonants, which is consistent with results

from previous experiments (Meyer, 2008; Meyer et al., 2017). We also observe that the average consonant categorization rate (at 55.1% for Expts 1A and 1B) is similar to that of whistled vowels (53.5%), even though participants' score in Expt 1B was 5.7% better compared to the categorization rate of isolated vowels in Expts 3 & 4.

In the context of the whistled word (Chpt.3.3), participants recognized the stimuli with an average of 45.6% of correct answers, well over chance at (20%). As this rate is lower than the isolated phoneme categorization rates (by 7.9% for vowels and by 9.5% for consonants), we suggested that the word context does not improve whistled speech perception for naive listeners, unlike in studies with native whistlers (see Chpt.3.3). We further investigated the role of individual phonemes in the word by comparing consonants and vowel correspondence (matching the phoneme played with the phoneme answered). We observed how consonants and vowels played different roles in word categorization according to position, where the vowels (according to their position) had a larger effect on word categorization than consonants.

6.1.2 Identifying whistled cues

Having established these high categorization rates, we then take an interest in the whistled cues used in this categorization process (thus addressing question 2).

By focusing on the whistled phoneme cues in Chapters 2 and 3, we first illustrate the importance of specific cues (highlighted by strong phonemic preferences) within the context of the isolated consonant, the isolated vowel and the vowel within the word, without considering musical experience.

6.1.2.1 Isolated whistled consonants

In the case of the "isolated" whistled consonants (tested in a VCV form with /a/), the results emphasize an advantage for pitch movements (or consonants with a "high locus"; Rialland, 2005) in contrast with consonants with no pitch movements - "grave" consonants (Chpt 2.1, Expt 1A, 2 and Chapter 2.2, Expt 1B). As we only tested four consonants, this advantage corresponded to an opposition between /t/ and /s/, "high locus" consonants which were recognized best (an average between both experiments of 71.65% and 74.12% respectively), and /k/ and /p/, "grave consonants" which were recognized with more difficulty (an average rate of 54.95% and 40.38% respectively). We suggest that the pitch change cue thus provides a first axis within which we differentiate these consonants. This underlines the importance of the parameter "acute"/ "grave" compared to that of the "continuous"/ "interrupted" cue (which would oppose the semi-continuous /s/ to the interrupted /k/, /p/ and /t/). Interestingly the important contrast between "acute" and "grave" consonants has also been shown in synthetic speech, where Liberman et al. (1954) opposed "plus shift" modulations (acute consonants) and "minus" transitions (grave consonants).

By comparing the whistled pitch ranges in Chpt 2.1 (high and low consonants) we further nuanced these consonant hierarchies. In Expt 1A, /s/ and /t/ show almost no differences in categorization rates (73.75% and 74.5%), however, in the lowered whistled consonant pitches (Expt 2), /s/ is categorized better than /t/ (82.5% compared to 41.25% in Expt 2). This is also the case in Expt 1B (74.5% compared to 68.8%) though there were no significant differences between the consonants. This suggests that among the two "acute" consonants tested, differences are observed in terms of the "continuous"/ "interrupted"

contrast. There were no differences among the grave consonants when comparing the two whistled ranges. However, in Expt 1B (Chpt.2.2), we observe significant differences among grave consonants, where /k/ is categorized better than /p/. However, this contrast does not apply in Expt 1A (Chpt.2.1), where /k/ shows no significant differences with /p/. In further specifying this difference in Chpt.2.1, we noticed an asymmetrical consonant confusion between /p/ and /s/, where /p/ was taken as /s/ 19.5% of the time, but when /s/ was heard, it was categorized as /p/ only 7% of the time. We suggested that the differences between /p/ and /k/, as well as the similarities between /p/ and /s/ may be due to differences in attack cues (according to amplitude or articulation), or to the "semi-continuity" described for /p/ in other studies (Diaz, 2008). The various contrasts between cues seem to suggest a hierarchical value of traits among consonants, with a strong dominance of the pitch change cue. Indeed, the high locus, or pitch change in /s/, seems more dominant than the "semi-continuity" or similar consonantal attack shared with /p/, given the large differences in categorization rates.

These findings nonetheless underline how categorization also relies on these secondary acoustic/articulatory parameters. We therefore can consider that the consonant categorization rates obtained reflect naive listeners' ability to use whistled cues such as continuity/interruption, articulation, and amplitude rise. Though the first two parameters (pitch change and interrupted/continuous) figure in various whistled consonant descriptions (see Chpt.1.2.2.1), the secondary cues suggested do not. We suggest that further exploration of these secondary cues would require the analysis of a larger range of consonants.

6.1.2.2 Whistled vowels

Among the whistled vowel results, we observe a very stable vowel hierarchy and cue preference in each of the vowel-focused chapters (i.e. both in isolation and in the context of

the word). We show that /i/ and /o/ are always categorized best, at around 86.04% and 58.95% respectively[17] (with an advantage for /i/ over /o/), and that /e/ and /a/ are more difficult to categorize, with rates around 43.56% and 38.31%[18]. Such a hierarchy corresponds to an advantage first for the highest pitched whistled vowel, followed by the vowel with the lowest pitch, with the middle vowels being the most difficult. This hierarchy echoes previous findings (Meyer et al., 2017) and confirms the robustness of this vowel representation: these vowel cues are well categorized regardless of the context and without being affected by inter- and intra-whistler differences. The consistency in results for the categorization of whistled vowels, due to consistent cues, reflects characteristics in modal speech. Indeed, when comparing Georgeton et al., 2012 and Gendrot & Adda-Decker, 2005, we observed how even though the vowel space became smaller due to context, the vowel distribution remained the same as in isolation. Inter- and intra-whistler differences also do not affect this hierarchy, which recalls findings in modal speech, notably concerning speaker normalization according to the 3-3.5 Bark difference (Strange, 1989; Syrdal & Gopal, 1986). We suggest that taking a closer look at the relationship between whistled speech and modal speech will help understand these categorization rates and the perceptual cues involved.

We can explain the advantages present for whistled /i/ and /o/ by considering these vowels as focal vowels. Indeed, the focal vowel /i/ is characterized (in French) by a convergence of F3 and F4 (high formant frequencies). As underlined by Meyer et al. (2017), the high center of gravity for /i/ in French may be coherent with the highest whistled pitch, thus reproducing underlying perceptual processes from modal speech. The center of gravity effect could also apply to /o/, especially when we consider that the range for /o/ in Silbo

[17] Statistics used from Expt 3
[18] Statistics used from Expt 3

overlaps with that of /u/ (see Chpt.1.2.2.2), a focal vowel characterized by the convergence of F1 and F2 (lower formant frequencies). Advantages for these vowels thus echo findings underlining the perceptual stability of focal vowels (Schwartz et al., 1998), as well as the universality of focal vowels. The difficulties presented for /e/ could also be due to a focal vowel bias, as /e/ is not a focal vowel. As some studies have shown that non-focal vowels are better categorized when moving away from the focal vowel, from /i/ to /e/ for example (Bohn & Polka, 2001), this relationship may help categorize /e/ over chance. Nonetheless, as /e/ is not a focal vowel, obtaining a lower categorization rate for /e/ is coherent with focalization theory (Zhao et al., 2019; see Massapalo et al., 2017; and Polka et al., 2021 for a review). However, the clear disadvantage for the focal vowel /a/ could contradict this hypothesis. Thus we wonder if, possibly due to differences in language, certain whistled vowels are a clearer manifestation of focal vowels than others. Indeed, Meyer et al. (2017) suggest that the confusions between the vowels /a/ and /o/ could be due to a different placement of these vowels in Spanish and French.

We further observed the role of vowel range through the inter-whistler variability. Indeed, the important role of the vowel range, or ambitus, constructed between the two extremities /i/ and /o/, was highlighted several times in our results. In Expt 3, we generally observed an advantage for the whistler with the wider vowel range (whistler B), which was confirmed in Expt 4 (60.23% of correct responses obtained for Whistler B, and 46.49% for Whistler A). We suggested that this was due to larger space between /i/ and /o/. The vowel-specific differences observed for the two whistlers also suggested that advantages for Whistler B were linked to this space. Indeed, participants showed an advantage for /i/, the highest vowel, produced by Whistler B in Part 3 rather than by Whistler A. Participants also showed an advantage for /a/ produced by Whistler A, (the whistler with a small vowel range)

in Part 3 over those of Whistler B. It is possible that the whistler with the larger range emphasizes the two vowel extremities (/i/ and /o/) first, allowing for a better performance on /i/ than with Whistler A productions. Whistler A (with the smaller range) may not function in the same manner, given the advantage shown for /a/.

The importance of frequencial distance between vowels was also observed for /o/ in the whistled word, where vowel correspondence rates varied according to position (see Chpt.3.3, Expt.5). In the V1 position, /a/ and /o/ were only 200 Hz apart (M: 1650.23 Hz and M: 1454.36 Hz respectively), and we observed no clear preferences for /o/ over /a/ or /e/. In the V2 position however, where /a/ was around 500 Hz higher than /a/ (M: 1638.60 Hz and 1137.91 Hz respectively), participants showed an advantage for /o/ over /a/ within the vowel correspondences. These findings also suggest that the construction of a wider whistled vowel range allows for a better representation of the vowels /i/ and /o/.

Overall, though these vowel hierarchies and the cue preferences highlight a strong capacity for whistled vowel categorization, the vowels analyzed are nevertheless characteristic of the phonological system of Canary Island Spanish. We wonder if a different choice of vowels, representing other aspects of the French vowel space, would affect participants' results differently.

6.1.3 The effect of musical experience

We can now consider the effect of musical experience on whistled speech perception by revisiting the three experimental structures used previously for isolated consonants, isolated vowels and whistled words. We consider these perceptual levels in three final questions:

(3) Will musical experience have an impact on whistled speech perception? (4) If so, which elements of the speech signal are affected? (5) How do these effects reflect specific skills acquired through musical training such as musical level achieved and instrument specialization?

6.1.3.1 Musical Advantage

Before comparing musician and non-musician participants, we can first underline how, in the experiments that include participants with musical experience (Chapters 4 and 5), we observe high categorization rates for phonemes, with an average of 69.1% correct answers for whistled consonants and 60.83% correct answers for whistled vowels (Expts 6 & 7). These phoneme categorization rates show a larger advantage for whistled consonants (9.17% higher than whistled vowels) than between Chapters 2 and 3. In Chapter 5, when including participants with musical experience, the whistled word was recognized with 57.4% correct answers. Thus, not only do participants with (and without) musical experience perform well over chance on whistled speech categorization at different perceptual levels, but groups which include participants with musical experience show higher categorization rates. The same relationships between vowel, consonant and word categorization observed in Chapters 2 and 3 are nonetheless maintained in Chapters 4 and 5. Indeed, like with isolated phonemes, we observe that phoneme correspondence rates in the word, at 64.1% for consonants and 68.4% for vowels (Expt 8), are higher than word recognition rates.

When considering and comparing the categorization rates for musicians and non-musicians on each perceptual level, we found a clear advantage for participants with musical experience. In the case of the isolated whistled vowel (Chpt.4.1), we observed an 8.2% difference between the average correct answers obtained by “musicians”, compared to the

average correct responses obtained by “non-musicians”. In the case of isolated whistled consonants (Chpt.4.2), we observed a 31.6% difference between high-level musicians and non-musicians, an 11.9% difference between the high-level musicians and the low-level musicians, and a 19.7% difference between the low-level musicians and the non-musicians. These differences for isolated phonemes suggest a larger advantage for high-level musicians over non-musicians when categorizing whistled consonants than in the whistled vowel categorization. Indeed, high-level musicians achieved higher scores on the consonant categorization task (77.5%, SD: 14.38) than on the vowel categorization task (64.6%, SD: 13.7), with a 12.9% difference in categorization rates.

This musical advantage is also observed in the context of the whistled word, notably in overall word categorization rates (Expt 8). We observed a 15.8% difference in categorization rates between participants with a high-level of musical experience and those with none, a 4.29% difference between high-level musicians and low-level musicians, and an 11.5% difference between participants with a low-level of musical experience and those with none. Contrary to the isolated phoneme tests, we observed slightly higher phoneme correspondence rates for whistled vowels than for whistled consonants, for both low-level musicians (with a 6% difference between vowels and consonants) and high-level musicians (with a 3.5% difference between vowels and consonants). These rates, thus oppose the preference for whistled consonants observed in the isolated categorization task. We suggest that this underlines a difference in behavior for phonemes within the context of the word compared to phonemes in isolation or VCV forms. However, this difference could also be attributed to the number of consonants included in the whistled word, which increases the number of consonant options.

More generally, these results highlight a clear effect of musical experience on performance, though these effects depend on the context and the task.

6.1.3.2 Cue-specific advantages

We can now further detail these results according to phonological preferences for vowels, consonants, and words, just as we did with the naive non-musician participants. This will help us identify how musical experience modifies the role of these cues.

For consonants in the VCV context (Chpt.4.2), hierarchies mirrored those observed in Chapter 2, where /s/ and /t/ were easiest to categorize and /k/ and /p/ were more difficult. High-level musician participants showed advantages over non-musicians/low-level musicians for the most difficult consonants /k/and /p/, as well as for the semi-continuous consonant /s/ (Chpt.4.2). For /k/, high-level musicians' results are significantly different from those of both low-level musician participants and non-musicians, while for /s/ and /p/ high-level musicians only show significant differences with non-musicians. Low-level musicians also show significant differences with non-musicians for /s/, indicating that even a little musical experience changes the way the cues for /s/ (acute and continuous) are treated. However, the advantages for /k/ and /p/ due to the improved use of the cues found in /k/ and /p/ (grave, interrupted – or secondary cues such as attack through amplitude or articulation) seems reserved for participants with high-levels of musical experience. Thus we suggest that increased musical competency changes the way different cues are treated. Interestingly, this does not affect consonant hierarchies, as all participants categorize whistled consonants similarly. It would therefore seem that high-level musician participants use the same tools and

cues as participants with less or no musical experience, but that their experience may allow them to analyze these secondary cues more finely.

We further specified the cues used by high-level musicians by focusing on instrument specialization (flute, piano, violin, and voice) for both consonants in isolation and in the word. This highlighted instrument-specific differences among these high-level musicians. Indeed, in the whistled consonants, all instrumentalists showed differences with non-musicians, but singers showed a learning effect, and flutists also showed significant differences with low-level musicians. This suggests that, depending on their instrument, musicians are able to exploit certain consonant cues rather than others. When considering the specific consonants, all instrumentalists showed advantages over non-musicians for /k/, violinists and flutists showed significant advantages for /s/ over non-musicians, and flutists also showed significant advantages for /p/ and /t/ over non-musicians. This shows a clear advantage for flutists.

In the whistled word, we also compared phoneme correspondence rates in order to observe consonant preferences, which we measured in comparison with expert whistlers. All instruments showed advantages for /p/ as there are no significant differences between their results and those of expert whistlers. This contrasts with the VCV context where /p/ was the least well categorized. Advantages (or similarities with expert whistlers) were also shown for /k/ by violinists, singers and flutists in the whistled word (we recall how this advantage was also shared by all instrumentalists in the VCV consonants). Finally we observed advantages for /t/ specific to flutists (like in the VCV form). For /s/ we observed significant differences with the expert whistlers for all instrumentalists (contrary to the advantages shown in the VCV form).

These results thus suggest that musicians show advantages in both consonant contexts for /k/ and /t/, whereas advantages for /s/ and /p/ are context specific. We suggest that the difficulty with acute consonants could be due to difficulties associating pitch changes (of various forms) to a single consonant (Chpt.5). In addition, integrating multiple pitches into one syllable may have affected categorization (as mentioned in the comments, see Annex A.5.1). We draw attention to this, as several participants indicated hearing more than two syllables in a word (see Annex A.5.1). Indeed, because the whistled pitch of "acute" consonants rises above that of the highest whistled vowel pitch (/i/), participants may have heard several vowels and therefore more syllables[19]. However in /t/ this ambiguity is clearly affected by the "interrupted" cue or the attack, as flutists categorize /t/ as well as expert whistlers.

In fact, the advantages for /k/ observed more generally, as well as for /t/ in the VCV form and /p/, suggest that high-level musician participants are able to hear and use the amplitude rise cue or articulation cues in these consonants. This may be due to their own experience with producing such forms of articulation in music, though articulatory parallels vary according to the instrument (flutists, for example, articulate /t/ and /k/ in order to articulate notes, whereas violinists use the bow and bow pressure). We could also attribute these advantages to the similarity in musical timbre between whistled speech and one's own instrument (well described by participants with musical experience – see Annex, A.4.1), as the timbre of the flute is, out of the instruments tested, the most similar to whistled speech. In other contexts, musicians have also been shown to rely on such cues, using timbre for example to process tonal languages (shown in MMN responses, see Hutka et al., 2015; Martínez-

[19] For example when the word /pase/ was played, it may have sounded like /pa.i.se/, see Annex A.2, Table 3.

Montes et al., 2013). This further suggests a hierarchy between cues within the consonants, where certain instrumentalists exploit finer cues according to instrument specialization.

In isolated vowels, we consistently observe the same vowel hierarchies for both musicians and non-musicians. We also observe a general effect of whistler B's wider vowel range on whistled vowel categorization for both musician and non-musician participants (see Expt 6), where results for whistler B were higher than those of whistler A for /i/, /a/, and /o/, thus emphasizing advantages in the extremities of the range (for /i/ and /o/), and possibly a better representation of intervals within that range.

When considering vowel-specific cues, though musician participants generally showed an advantage for /a/ over non-musicians (20.5% difference), they were also strongly impacted by the whistler heard. Indeed, musical advantages were present for the middle vowel /e/ and the lowest vowel /o/ only for whistler B (the whistler with a larger range), (cf. Chpt. 4.1). There were no vowel-specific musical advantages for whistler A. The advantages observed within whistler B may suggest that musicians are able to use the vowel space created with a wider whistled range to their advantage. Here, these advantages underline changes in behavior for the lower or middle vowels, and we wonder if this may reflect an awareness of the vowel organization within the whistled range. This is suggested in some of the participant comments, where one musician mentions using the relationship between whistles by comparing them to a chord (see Annex, A.5.1).

We also observe similarities between the expert whistlers and the high-level instrumentalists for certain vowels in the context of the whistled word. Indeed, violinists, flutists and singers (cf. Chpt 5) show advantages for /o/ and /i/, where they perform similarly to whistlers. This was not the case for /a/ or /e/, where whistlers perform better than

musicians. This highlights improved processing mechanisms for the extremities of the whistled range rather than for the middle vowels. Thus, though participants with high-levels of musical experience perform significantly better compared to non-musicians, these advantages are lesser than those of whistlers, and not as present for whistled vowels. This suggests that the transfer of skills between participants with high-level musical experience may not improve perception in the same way as experience with whistled speech does, as whistled knowledge may be more relevant (like in Simpson, 1975).

6.2 Discussing whistled speech perception

6.2.1 Understanding the perceptive process

6.2.1.1 Whistled speech and perception theories

As highlighted in the previous sections, these findings allow us to better understand how whistled speech is perceived and how musical experience impacts categorization tasks on phonemes and words. These findings thus allow us to discuss the speech perception process more generally, by taking an interest in the acoustic and articulatory cues employed for categorization and their relationship with established theories.

In the whistled consonant perception tasks, we highlighted how pitch-change dominated over the "interrupted"/"continuous" cue. Superficially, these two cues characterize phonemes acoustically, and the clear changes in the acoustic signal are easily compared to musical compositions (low/high note choices, and adding rests between notes). However, we observed the presence of secondary cues, distinguishing /k/ and /p/ for example, in both the non-musician and musician participants' performances. Acoustically, we proposed that /k/ and /p/ could be distinguished if /p/ was considered as "semi-continuous", though this was not the case in our corpus (see Annexe, A.2. Table 2) or by comparing consonant attacks. Though this attack can also be characterized acoustically, through the amplitude rise, the differences in amplitude rise between /k/ and /p/ are not always clear (see for example productions 2 and 4 in A.2.Table 2). This leaves elements such as articulation point (velar, bilabial), or more subtle acoustic cues such as stop duration (see for example Ridouane, 2018) which were not tested here. Thus, though we can characterize whistled consonant cues through acoustics, these give way to secondary articulatory elements. In addition, the initial

acoustic cues such as pitch change and interruption are inherently linked with articulation – highlighting the role of both articulatory and acoustic elements in whistled consonants perception.

In whistled vowels, our results highlight preferences for the highest and lowest vowel pitches, /i/ and /o/, where /a/ and /e/ were more difficult to categorize. These simplified vowels are in their essence characterized by acoustic cues, through pitch ranges specific to each vowel (and specific to the whistler and elements of co-articulation). The general instability of each vowel (see for example Chpt.3.1, Figure 2 or Chpt.3.3, Figure 2) indicates that categorization cannot be based entirely on acoustic cues, but must depend on relationships between vowels (as shown through inter-whistler variability). These relationships can be described through a mathematical formula (see Chpt.3.1), however the placement of each of these pitches within the vowel space depends on the articulatory properties. These properties then affect the size of the vocal cavity and the formants present. Thus, just as for the whistled consonants, we observe a complementary role between both the acoustic cues and articulatory cues.

How do these findings reflect speech perception models? As described previously, the Auditory theory proposed by Fant (1960) relies on invariable acoustic cues observed in the speech signal. These acoustic cues lead to phonemic representations, which are then used for speech production. In some of our results, we observe the presence of some invariable acoustic cues: for example pitch change for /t/ and /s/, or the order of pitch in the vowels tested, which correspond to formant shapes in modal speech. However, we also observed how elements of variability stemming from co-articulation or inter-whistler differences have a clear effect on categorization based entirely on whistled cues. Thus, in the reduced form of

whistled speech where acoustic redundancy is eliminated, this variability strongly affects the role of these acoustic cues. This suggests that participants also rely on certain articulatory elements (either as secondary cues or underlying causes for these acoustic elements) for categorization.

In Motor Theory (Liberman & Mattingly, 1985) the role of articulatory cues comes to a forefront, where the cues observed are not acoustic, but articulatory – either visually, or through invariant phonetic gestures. This cannot entirely apply to whistled speech, due to the production methods which impede both visual tools as well as some of the articulatory mechanisms used in modal speech. Motor Theory then automatically links the perception of these phonetic gestures with knowledge of production. Though this may be applicable for expert whistlers, naive listeners have never used whistled speech and cannot apply their knowledge of production to perception. Indeed, this highlights how for naive listeners hearing whistled speech, acoustic cues are surely prioritized, as they adapt to the modified speech cues.

Thus we propose considering models which integrate both acoustic and articulatory cues in speech perception. One such model is the PACT Theory (Schwartz et al., 2012). In this model, gestures serve as perceptuo-motor units, shaped and selected by perception according to their "audibility". Thus, in the PACT Theory, perception and production are linked, where relevant acoustic cues and articulatory cues are employed for categorization. This theory, though one among many, provides an interesting perspective on whistled speech perception, through the lense of the naive listener's experience. Indeed, these listeners may not know how to create the whistled gestures heard, however they nonetheless use the cues present based on their own inherent knowledge of phonemes – which include both acoustic and

articulatory elements. Such perspectives provide insight into speech perception more generally.

6.2.1.2 Perceptual levels

In our discussion of the speech perception theories, we focused more specifically on phoneme categorization. However, in our studies, we also investigated the relationship between whistled phoneme perception and whistled word perception.

When testing words, we observed large differences in categorization rates depending on the vowels and consonants present in the word. This suggests that word categorization is influenced by its internal phonemes.

Interestingly, we notice a difference in effect between the vowel and the consonant within the word, possibly reflecting the role of these cues as prelexical units. Indeed, the stability of the whistled vowel categorization rates and hierarchies (in isolation, in the word, and through vowel hierarchies) suggests that vowels play a different role in the word than consonants, similarly to modal speech (cf.Chpt.3.3)[20]. Consonant productions, however, were heavily impacted by the co-articulation present in the word, affecting cue production and consonant hierarchies. Interestingly, this is similar to findings with synthetic speech, where consonants coarticulated with /i/ were more difficult to categorize than those with /a/ (Liberman et al., 1954). Thus, in the context of whistled speech, we wonder if the consonant cues are more relevant to whistled word perception than the phonemes themselves. This would generally mirror the way Rialland (2005) used whistled cues to test for consonant

[20] However, we note that the distinction between consonants and vowels (lexical function/grammatical function) as proposed by Bonatti et al. (2005) is difficult to test or confirm as our word categorization task was very limited.

groups[21]; that is by applying perceptual tests to the “acute”/”grave” contrast (coronal and non-coronal), and the “interrupted”/”continuous[22]” contrast (voiced/unvoiced). Constructing such cue-based groups could also allow us to extend these analyses to other consonants in these studies that share similar cues. Indeed, we wonder how participants distinguish consonants that share the same whistled cues (for example /s/ and /ʃ/).

The effect of phonemes on word perception allows us to consider word-perception models commonly used to describe this process. In our studies, we proposed a shortened word list for participants to choose from, thus eliminating a strong effect of top-down perception or feed-backwards. However, we nonetheless observed an effect of the word on phoneme categorization and therefore an interaction between the phoneme and the word (used in models such as TRACE; McClelland et al., 1986). This was suggested through the effect of co-articulation, notably with the consonant /s/ where naive listeners had much greater difficulties with words which included /s/, contrary to their performance on VCV syllables. We suggest that this could be due to two effects of co-articulation: the lack of pitch-change cues in the C1 position or the lengthening of syllables due to the continuous (non-interrupted) pitch-changes of /s/ in the C2 position. The reverse applied to /p/, which was very difficult to categorize in the VCV form, but benefited from the context of the word. This suggests that the word therefore interacts with phoneme perception and categorization.

In discussing the interaction between prelexical units, we also wonder how whistled speech cues compare with the prelexical units proposed in the word models reviewed (see Chpt.1.1.1.1). In models such as Cohort (Marslen-Wilson & Welsh, 1978), DCM (Gaskell &

[21] In these tests, the consonants tested were all of the form VCV, and not in the word. Therefore the variability of cues according to position was irrelevant.
[22] Though we note that in this test /s, k, p, t/ are all interrupted consonants

Marslen-Wilson, 1997), and TRACE, whistled cues (articulatory/acoustic) could easily serve as "features", though, as suggest by Rialland (2005) these cues almost serve as phoneme groups. However, when considering models which rely on phonemes, it seems unlikely that the whistled phoneme would serve as a clearly defined unit which can serve as a basis with which to create "chunks" (ART model; Magnussen et al., 2012), or be correctly added/subtracted (NAM; Luce & Pisoni, 1998), due to ambiguities (for example with /s/). This ambiguity may even suggest that the syllable is a relevant unit in this categorization process (Massaro, 1972; Mehler et al., 1981; Mattys & Melhorn, 2005), perhaps serving as an intermediate perceptual level used to deconstruct the word.

6.2.2 Musical transfer and musical skill

6.2.2.1 Types of musical transfer

When considering the second group of questions on musical experience (3, 4, 5), we quickly confirmed the transfer of skills from music to speech for participants with musical experience. This allows us to then focus on the type of musical transfer applied by considering which elements of speech are most affected.

We observed an effect of musical experience on both isolated vowels and VCV consonants, however, we also notice a larger advantage in the whistled consonants than in the whistled vowels. The specificity of the musical advantage is further demonstrated in advantages for specific consonants and vowels and in certain contexts (for whistler B for example, or for /s/ in the VCV form, but not the word). These advantages suggest that musical experience does not affect all aspects of speech perception, transferring only to elements

which are similar to their musical training. As described in Chpt.4.1, musical training improves various aspects of perception – perceptual (acoustic) acuity for example, may allow musical participants to observe the secondary cues present in /k/ or /p/. Pitch matching, or relative pitch awareness may affect vowel placement within the vowel space. These skills interact, as observed with whistler B, where clearer acoustic distances (therefore requiring less minute perceptual acuity) may improve relative pitch awareness.

We further observed the impact of specificity in musical experience when focusing on instrument specialization. This showed an effect for both consonants (Chpt.4.2 and 5.) and vowels (in the word, Chpt.5.1, and in Annexe A.4.2), where individual instrumentalists performed better on specific phonemes. This suggests that even though musical experience only affects elements of whistled speech where transfers can occur, the type of musical experience received further defines these advantages. By using instrumental experience, we opposed "general" musical advantages which are common to all musicians, such as cognitive skills, attention and memory, with instrument specific skills, which are specific to the sound and production techniques of their instrument. As we observed differences in instrument specialization, we suggested that these "musical advantages" hint at a transfer based on sound-specific knowledge. Reprising Barbaroux (2019), these findings correspond to a "waterfall" model rather than a "multidimensional" model (see Chpt.4.2), suggesting that cognitive skills such as memory and attention do not seem to be key to such musical advantages. Thus we suggest that in the context of whistled speech, only near-transfers take place. As such sound-specific improvements have been previously observed in other forms of modified speech (see Bidelman & Krishnan, 2010; Varnet et al., 2015), it is not surprising to suggest that they also apply to whistled speech. Furthermore, we highlight how the transfer of musical knowledge towards speech perception is amplified by similarities between the form

of speech perceived and the specific musical knowledge of the participants, as it provides a better representation of the stimuli (see Parbery-Clark et al., 2009). This holds true in the transfers observed in whistled speech concerning musical timbre and articulation (see Chpt.4.2).

The effect of musical experience on perceiving whistled speech is also specific to the amount of musical skill acquired. Indeed, in each of the studies including musically experienced participants, we observe larger advantages for high-level musicians than for lower-level musicians. Yet among the high-level musicians (levels 4, 5, and 6), we also observed very few differences in categorization rates. For example, in the VCV consonant categorization task in Chapter 4.2, participants showed only a 2.6% difference between levels 4 and 5, and a -1.7% difference between levels 5 and 6. This suggests that for the VCV consonant, after reaching a certain musical level, the transfer is relatively homogeneous (i.e. there is no difference between levels within the high-level musicians). This may not be the case, however, for the whistled word, where differences in whistled word categorization rates among high-level musician participants differ by 8.55% between levels 4 and 5, -12% between levels 5 and 6, and only by 1% between levels 3 and 4 (see Chpt.5, Figure 2). This highlights a change in profile among high-level musicians, with a marked decrease in scores for professional musicians. We also observed that participants in level 3 perform very similarly to those in level 4. These specificities suggest that while a high-level of musical experience gives participants advantages in whistled speech perception, the transfer is not homogeneous. These fluctuations in performance (decreasing or plateauing), could indicate that skill-sets required for certain tasks are achieved with a lower level of musical experience, or that very high-levels of musical experience negatively impact transfers between music and speech.

Finally, though we observed knowledge transfers for participants with musical experience towards whistled speech, we underline that this musical advantage is not as important as a transfer due to training in whistled speech (see Chpt.5). Indeed, in Chapter 5, experienced whistlers (even when listening to whistled words of a different language), show clear advantages in the task compared to musician participants. This difference is coherent with the advantages shown by Spanish whistlers compared to naive listeners when tested for isolated vowel categorization in Meyer (2008). We suggest that these advantages are due to perceptual acuity, and knowledge of whistled speech production, as the words heard were produced in a different language, and participants would not have been able to rely on lexical activation for these improvements. Thus, as our studies included small amounts of perceptual training that showed no (or very little) effects on naive listeners' categorization, we wonder if training on production may have a larger impact on whistled speech perception. Indeed, we suggested that this may be the key to flutists' advantages, and if so, further investigations into the role of production in whistled speech perception could be fruitful.

6.2.2.2 Perceptual Processing

The findings observed in each of the studies presented also allow us to reconsider music and speech processing, and their interaction. Peretz & Coltheart (2003), propose an analysis of sound processing where speech and music are first treated with a common analysis, before transitioning towards a specialized music module, language module, or a still-not-characterized module. In our studies, we observe a transfer between music and speech which is coherent with studies showing shared processing mechanisms and common resources, which can cause interference between the two (Atherton et al., 2018). When applying the Peretz & Coltheart (2003) model here, whistled speech should be considered as

part of the “speech” module (as indications are given declaring the stimuli as speech in our experiments). However, in order for transfers or interference to occur between musical processing and speech processing, we suggest that the two processing modules must be overlapping in order to share aspects which are specific to each module.

In order for the overlap between the music module and the speech module to exist, listeners must have a sufficient amount of musical expertise to use that knowledge as a tool for speech processing. In the case of whistled speech, this seems to apply only to high-level musicians. The musical tools which transfer towards the speech model also only apply to elements of speech which are shared with music.

This shared processing mechanism represented through the overlap between modules may also have limitations. Indeed, the interference observed by Atherton et al. (2018) is an example of this: instead of borrowing information from the music module, the overlap between modules interferes with speech (or music) processing. In our results, we observed a decline in the results of the professional musicians compared to participants with musical levels at 4 or 5. In this case, we suggest that due to their highly established musical specialization, there is no overlap between modules. Thus, Figure 3 would only represent perceptual processing for participants with a certain amount of musical experience, enough to reapply musical skills, but who are not overly specialized in the field, where musical tools apply only to music. More generally, this perspective highlights the presence of transfers between music and speech due to shared or similar processing mechanisms present in each module. Thus we suggest that this invalidates the idea of specificity of speech perception as a unique process, as described in the initial theories of Motor Theory (Liberman et al., 1967;

Liberman & Mattingly, 1985), as music is treated similarly enough to transfer and interfere with speech perception.

6.2.2.3 Qualifying musical experience for future studies

Throughout the studies described here, we have proposed various different evaluations of musical experience, both through self-declared skill levels and through instrument specialization. In doing so we hoped to address the difficulty in defining the "musician" and provide a more detailed understanding of how specific musical skills affect speech perception. Indeed, in previous experimental studies targeting speech perception, the qualification of "musicians" varied greatly in terms of their musical experience, often including participants with various skill levels, who play contrasting genres of music, and different instruments (cf. Table 1, Annex, A.1). To avoid such disparity, we focused on classical music, defining the "musician" according to the French Conservatory system, while targeting high-level musicians. We then divided participants in several groups: in Experiment 6, we used a binary opposition between "musicians" and "non-musicians". However, in a recent review by Smit et al. (2023), they suggested the importance of a continuous qualification of musical ability in experimentation, instead of opposing musician and non-musician groups. In Experiments 7 and 8, we therefore divided participants into three groups (non-musicians, low-level musicians, and high-level musicians). In doing so, we observed increasing advantages according to musical level, with only a slight decrease for level 6 in the whistled word. Further specifications according to musical instruments have also enabled us to better define musical transfers, and explore the relevance of cues present in the signal.

Thus, these results underline the importance of defining the musician in detail according to their musical training level (and skill sets), as well as instrument specialization in

order to best understand the effect of musical experience on speech, or in transfers more generally. However, despite our best efforts to target musician participants in a concise way, there were nonetheless a number of setbacks in the way musical skills were considered here. We hope that a final discussion of these issues will provide sound and clear suggestions for future studies.

The first of the issues present in our definition of musical experience is the self-declared musical level. Recent studies using the "Musical Sophistication Index" (either the Ollen MSI or the Goldsmith MSI) have shown that the single item measure "what title best describes you" (in terms of musical identity) corresponds closely to the number of years of private lessons (Zhang & Schubert, 2019). However the number of years of experience is not necessarily representative of the skills obtained. Relying on the self-declared musical level is similar to the single-item measure proposed by the Musical Sophistication Index, as it remains subjective and broad. Indeed, one's perception of level depends on the teacher, the student, the instrument and the environment. Because of this heterogeneity, participants without musical diplomas (levels 1, 2 and 3 in the proposed scale) may have certain musical skills equivalent to participants with musical diplomas (levels 4, 5 or 6). It is also possible that participants who qualified their skills as "amateur" (level 2) were better at certain musical tasks than participants who qualified their musical skills as "confirmed" (level 3). Therefore this suggests that for participants with musical experience who do not have diplomas, i.e. participants with "low-level" musical experience, testing skills may be more appropriate in order to understand musical competencies. This could include using a questionnaire with skill-related questions or tests, such as Goldsmith's Musical Sophistication Index, or testing for a specific skill (such as absolute pitch, also shown to be measured on a continuum, Leite et al., 2016). Such tests are however unlikely to consider elements such as instrumental skills, or

musicality, which are essential for musicians. For higher-level musicians (for whom we observed more significant musical advantages) diplomas include these elements – both in terms of defined technical skills and skills such as musicality which are more difficult to test. Diplomas therefore seem to serve as an appropriate measure for high-level musicians[23], but make skill comparisons with low-level musicians difficult.

The complexity of musical skill is also present when considering instrument specialization and musical genre. Indeed, many high-level musicians play several (often complementary) instruments at a high level (some examples include singers who play the piano, or pianists who play the harpsichord). In light of this multi-instrument specialization, it may be difficult to fully target one specific instrument-based knowledge, or to do so, one would require a large pool of participants. In addition, differences in instruments can also correspond to playing several musical styles. Though this study focused on classical music expertise, Tervaniemi et al. (2015) have shown that musicians who play different styles attribute different levels of importance to sound features, thus affecting the way they listen.

These details, applicable more specifically to the studies described in this thesis, highlight the complexity of defining musical experience. We can attribute some of these difficulties to the sheer number of years required to develop musical skills, making it impossible to conduct long-term longitudinal music studies for example (see Olszewksa et al., 2021). Our suggestions nonetheless attempt to improve future qualifications of musical and instrumental skill in speech perception experiments. They seek to provide a baseline for future investigations which can better consider the complexity of the musician.

[23] This pertains more specifically to classical music, as some other musical styles may not offer diplomas justifying one's musical level (see Chpt.1.3.3.2).

6.3 Perspectives

In this discussion, we responded to the initial questions proposed in this thesis, before considering speech perception and musical transfers more generally.

We first demonstrated that naive listeners can categorize and recognize whistled speech correctly in the isolated phoneme (consonants and vowels) and in the context of the word. These findings are promising, as they suggest whistled speech is a good tool for understanding speech perception and could be used in future experiments.

We then observed phoneme-specific categorization rates, which highlighted several essential cues used for categorization (pitch-change, or pitch contrasts, continuity/interruption…). These cues remained constant despite incorporating inter and intra-variability in productions. Thus, by using these reduced whistled versions of spoken phonemes, we approach speech perception and cue invariance in the context of natural production. Indeed, many studies using modified speech acoustically modify aspects of the signal (adding noise for example, modifying pitch cues, or creating synthetic speech), and thereby remove elements of natural variation present in production. Our studies highlight how natural variability allows us to consider speech perception models, and suggest that integrating elements of variability may provide insight into perceptive processes.

Finally, we observe a strong effect of musical experience on whistled speech perception. We confirm the possibility for transfers between music and speech, and detail how including variability within musical experience also provides a better understanding of the effect of musical skill, and the elements affect by musical skills. These findings suggest that

the role of secondary cues in speech are more easily exploited by musician participants, whose behavior varies according to the level achied, and the instrument played.

Thus, these findings provide the first steps on a path towards understanding speech perception from the perspective of speech variability. In this thesis, I show how using various forms of variability in the speaker and the listener (here whistled speech and musical experience) provides insight into speech perception processes. In our studies, we consider speech cues in various perceptual levels, apply speech perception theories on whistled speech, and investigate knowledge transfers stemming from the listener. In doing so, we use the complexity of both the speaker and the listener to our advantage, all while providing a more realistic representation of the perceptive process that is speech.

Bibliography

About the Suzuki Method. (2022). Suzuki Association of the Americas. Retrieved September 27, 2022, from https://suzukiassociation.org/about/suzuki-method/

Adler, S. (2002). *The Study of Orchestration*. Norton & Company, New York.

Alexander, J. A., Wong, P. C., & Bradlow, A. R. (2005). "Lexical tone perception in musicians and non-musicians," in *Ninth European Conference on Speech Communication and Technology* (Lisbon).

Alvar, M. (1955). Las hablas meridionales de España y su interés para la lingüística comparada. *Revista de Filología Española*, *39*(1/4). https://doi.org/10.3989/rfe.1955.v39.i1/4.1136

Alwan, A., Jiang, J., & Chen, W. (2011). Perception of place of articulation for plosives and fricatives in noise. *Speech Communication*, *53*(2), 195.

Anvari, S. H., Trainor, L. J., Woodside, J., & Levy, B. A. (2002). Relations among musical skills, phonological processing, and early reading ability in preschool children. *Journal of Experimental Child Psychology*, *83*(2), 111–130. https://doi.org/10.1016/s0022-0965(02)00124-8

Atherton, R. P., Chrobak, Q. M., Rauscher, F. H., Karst, A. T., Hanson, M. D., Steinert, S. W., & Bowe, K. L. (2018). Shared Processing of Language and Music. *Experimental Psychology*, *65*(1), 40–48. https://doi.org/10.1027/1618-3169/a000388

Bangert, M. & Schlaug G. (2006). Specialization of the specialized in features of external human brain morphology. *Eur J Neurosci*, 24(6), 1832-1834. https://doi.org/10.1111/j.1460-9568.2006.05031.x

Barbaroux, M. (2019). *Pratique musicale et effets de transfert : de la perception à la cognition*. Thèse. Université d'Aix Marseille.

Baudouin de Courtenay, J. (1972). *The Beginnings of Structural Linguistics*, A Baudouin de Courtenay Anthology, tr. and ed. by E. Stankiewicz. Bloomington: Indiana University Press

Benki, J.R. (2003). Analysis of English Nonsense Syllable Recognition in Noise. *Phonetica* 60, 129–157.

Bent, T., Kewley-Port, D. & Fergusson, S. (2010). Across-talker effects on non-native listeners' vowel perception in noise, *Journal of the Acoustical Society of America*, vol. 128, no.5, 3142–3151.

Bermudez, P., Lerch, J. P., Evans, A. C. & Zatorre, R. J. (2009). Neuroanatomical correlates of musicianship as revealed by cortical thickness and voxel-based morphometry. *Cereb. Cortex* 19, 1583–1596. https://doi.org/10.1093/cercor/bhn196

Bernstein, L. (1973). *The Unanswered question: Six Talks at Harvard*. Harvard University Press.

Besson, M., Schön, D., Moreno, S., Santos, A. & Magne, C. (2007). Influence of musical expertise and musical training on pitch processing in music and language. *Restorative Neurology and Neuroscience*, 25(3-4), 399-410.

Besson, M., Chobert, J., & Marie, C. (2011). Transfer of Training between Music and Speech: Common Processing, Attention, and Memory. *Frontiers in Psychology*, *2*. https://www.frontiersin.org/articles/10.3389/fpsyg.2011.00094

Best, C. (1995). A direct realist view of crosslanugage speech perception. In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, 171–204.

Bhide, A., Power, A., & Goswami, U. (2013). A Rhythmic Musical Intervention for Poor Readers: A Comparison of Efficacy With a Letter-Based Intervention. *Mind, Brain, and Education*, *7*(2), 113–123. https://doi.org/10.1111/mbe.12016

Bidelman, G. M., & Krishnan, A. (2010). Effects of reverberation on brainstem representation of speech in musicians and non-musicians. *Brain Research, 1355,* 112–125. https://doi.org/10.1016/j.brainres.2010.07.100

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch. *Brain and Cognition*, *77*(1), 1–10. https://doi.org/10.1016/j.bandc.2011.07.006

Bidelman, G. M., Hutka, S., & Moreno, S. (2013). Tone Language Speakers and Musicians Share Enhanced Perceptual and Cognitive Abilities for Musical Pitch: Evidence for Bidirectionality between the Domains of Language and Music. *PLOS ONE*, *8*(4), e60676. https://doi.org/10.1371/journal.pone.0060676

Bigand, E., & Tillmann, B. (2022). Near and far transfer: Is music special? *Memory & Cognition*, *50*(2), 339–347. https://doi.org/10.3758/s13421-021-01226-6

Blanco, N., Meyer, J., Hoen, M. and Meunier, F. (2018). Phoneme resistance and Phoneme Confusion in Noise: Impact of Dyslexia, *Interspeech*, 2290-2294. https://doi.org/10.21437/interspeech.2018-1271

Bonatti, L., Peña, M., Nespor, M. & Mehler, J. (2005). Linguistic Constraints on Statistical Computations: The Role of Consonants and Vowels in Continuous Speech Processing. *Psychological Science,* 16 (6), 451- 459. https://doi.org/10.1111/j.0956-7976.2005.01556.x

Bowers, J. S., Kazanina, N., & Andermane, N. (2016). Spoken word identification involves accessing position invariant phoneme representations. *Journal of Memory and Language*, *87*, 71–83. https://doi.org/10.1016/j.jml.2015.11.002

Bresin, R., & Battel, G.U. (2000). Articulation Strategies in Expressive Piano Performance Analysis of Legato, Staccato, and Repeated Notes in Performances of the Andante Movement of Mozart's Sonata in G Major (K 545). *Journal of New Music Research*, *29*(3), 211–224. https://doi.org/10.1076/jnmr.29.3.211.3092

Brós, K., Zygis, M., Sikorski, A. & Wołłejko, J. (2021). Phonological contrasts and gradient effects in ongoing lenition in the Spanish of Gran Canaria. *Phonology*, 38, 1-40. https://doi.org/10.1017/S0952675721000038

Busnel R-G., Moles A. & Vallencien B. (1962). Sur l'aspect phonétique d'une langue sifflée dans les Pyrénées françaises. *Proceedings of the International Congress of Phonetical Science, Helsinki*, 533-546.

Busnel, R-G. (1968). *Etude radiocinématographique d'un siffleur turc de Kusköy*. Paris : Service du Film de Recherche Scientifique.

Busnel, R-G. (1970). Recherches expérimentales sur la langue sifflée de Kusköy. *Revue de Phonétique Appliquée*, 14/15, 41-57.

Busnel, R-G. & Classe, A. (1976). *Whistled languages*. Springer-Verlag. Berlin Heideleberg.

Byun, T. M., & Tiede, M. (2017). Perception-production relations in later development of American English rhotics. *PLOS ONE*, *12*(2), e0172022. https://doi.org/10.1371/journal.pone.0172022

Calliope. (1989). *La Parole et Son Traitement Automatique*. Paris: Masson.

Carey, D., Rosen, S., Krishnan, S., Pearce, M. T., Shepherd, A., Aydelott, J. & Dick, F. (2015). Generality and specificity in the effects of musical expertise on perception and cognition. *Cognition*, *137*, 81–105. https://doi.org/10.1016/j.cognition.2014.12.005

Carlson R., Fant G. & Granström B. (1974). Two-formant models pitch and vowel perception. *Auditory analysis and perception of speech.* 55–82.

Casserly, E. D., & Pisoni, D. B. (2010). Speech perception and production. *Wiley Interdisciplinary Reviews. Cognitive Science*, *1*(5), 629–647. https://doi.org/10.1002/wcs.63

Chan, A.S., Ho, Y. & Cheung, M. (1998) Music training improves verbal memory. *Nature* 389, 128. https://doi.org/10.1038/24075

Chang, S., Plauché, M. C., & Ohala, J. J. (2001) Markedness and consonant confusion asymmetries. In E. Hume & K. Johnson (eds.). *The role of speech perception in phonology*. San Diego CA: Academic Press. 79-101.

Chiba, T. & Kajiyama, M. (1958). *The vowel, its nature and structure*. Tokyo, Phonetic Society of Japan.

Chistovich, L.A. & Lublinskaya, (1979). The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research*, 1 (3), 185-195. https://doi.org/10.1016/0378-5955(79)90012-1

Clark, J.F., Dermody, P. & Palethorpe, S. (1985). Cue enhancement by stimulus repetition: Natural and synthetic speech comparisons. *Journal of the Acoustical Society of America*, 78, 458–462.

Classe, A. (1956). Phonetics of the Silbo Gomero. *Archivum linguisticum*, 9, 44-61.

Classe, A. (1957). The whistled language of La Gomera. *Scientific American*, 196: 111-124.

Cohen, A. J. (2019). Singing. In P. J. Rentfrow & D. J. Levitin (Eds.), *Foundations in music psychology: Theory and research*, 685–750.

Cooper, F. S., Delattre, P.C., Liberman, A.M., Borst, J.M. & Gerstman, L.J. (1952). Some experiments on the perception of synthetic speech sounds. *J acoust Soc Am*, 24, 597–606.

Cooper, P. K. (2020). It's all in your head: A meta-analysis on the effects of music training on cognitive measures in school children. *International Journal of Music Education*. 38 (3), 321-336.

Coumel, M., Christiner, M., & Reiterer, S. M. (2019). Second language accent faking ability depends on musical abilities, not on working memory. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00257

Cowan, G. (1948), *Mazateco Whistled Speech*. Linguistic Society of America.

Cutting, J. & Day, R. (1975). The Perception of Stop-liquid clusters in phonological fusion, *Journal of Phonetics*, 3, 99-113.

Dahan, D., & Mead, R. L. (2010). Context-conditioned generalization in adaptation to distorted speech. *Journal of Experimental Psychology. Human Perception and Performance*, *36*(3), 704–728. https://doi.org/10.1037/a0017449

Danielsen, A., Nymoen, K., Langerød, M. T., Jacobsen, E., Johansson, M., & London, J. (2022). Sounds familiar(?): Expertise with specific musical genres modulates timing perception and micro-level synchronization to auditory stimuli. *Attention, Perception, & Psychophysics*, *84*(2), 599–615. https://doi.org/10.3758/s13414-021-02393-z

Davis, M. H., and Johnsrude, I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. Hear. Res. 229, 132–147. https://doi.org/10.1016/j.heares.2007.01.014

Defilippi, A. C. N., Garcia, R. B., & Galera, C. (2019). Irrelevant sound interference on phonological and tonal working memory in musicians and nonmusicians. *Psicologia, Reflexao e Critica: Revista Semestral Do Departamento de Psicologia Da UFRGS*, *32*(1), 2. https://doi.org/10.1186/s41155-018-0114-z

Degé, F., Kubicek, C. & Schwarzer, G. (2011). Music lessons and intelligence: A relation mediated by executive functions. *Music Perception 29*(2), 195–201.

Degé, F. (2021). Music Lessons and Cognitive Abilities in Children: How Far Transfer Could Be Possible*. Frontiers in Psychology,* 11. https://doi.org/10.3389/fpsyg.2020.557807

Delattre, P. C., Liberman, A.M., Cooper, F. S., and Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel colour: Observationosn one-and two-formant vowels synthesized from spectrographic patterns. *Word* 8, 195-210.

Delle Luche, C., Poltrock, S., Goslin, J., New, B., Floccia, C. & Nazz, T. (2014). Differential processing of consonants and vowels in auditory modality: A cross-linguistic study. *Journal of Memory and Language*, 1-15.

Delogu, F., Lampis, G., & Olivetti Belardinelli, M. (2006). Music-to-language transfer effect: May melodic ability improve learning of tonal languages by native nontonal speakers? *Cognitive Processing*, *7*(3), 203–207. https://doi.org/10.1007/s10339-006-0146-7

Díaz, D. (2017) (2008). *El lenguaje silbado en la Isla de El Hierro* (segunda edicion ampliada), Tenerife: Le Canarien ediciones, La Orotava.

Dickey, B. and Lasocki, B. (2020). *Tonguing*, Grove Music Dictionary Online*,* ed. D. Root.

Dittinger, E., Barbaroux, M., D'Imperio, M., Jäncke, L., Elmer, S., & Besson, M. (2016). Professional Music Training and Novel Word Learning: From Faster Semantic Encoding to Longer-lasting Word Representations. *Journal of Cognitive Neuroscience*, *28*(10), 1584–1602. https://doi.org/10.1162/jocn_a_00997

Dommelen, W.A. & Hazan, V. (2012). "Impact of talker variability on word recognition in non-native listeners". *The Journal of the Acoustical Society of America*. vol. 132, no.3, 1690-1699.

Duffy, S.A. & Pisoni, D.B. (1992) Comprehension of synthetic speech produced by rule: a review and theoretical interpretation. *Lang Speech*. Oct-Dec, 35 (Pt 4), 351-89. https://doi.org/10.1177/002383099203500401

Durand, P. (1985). Variabilite Acoustique Et Invariance En Français : Consonnes occlusives et voyelles. *CNRS Editions Du Centre National De La Recherche Scientifique*, Paris

D'Souza, A. A., Moradzadeh, L., & Wiseheart, M. (2018). Musical training, bilingualism, and executive function: Working memory and inhibitory control. *Cognitive Research: Principles and Implications*, *3*(1), 11. https://doi.org/10.1186/s41235-018-0095-6

Eccles, R. van der Linde, J. M., Holloway, J., MacCutcheon, D., Ljung, R. & Swanepoel, D. (2020) Effect of music instruction on phonological awareness and early literacy skills of five- to seven-year-old children. *Early Child Development and Care*, 191, 12, 1896 – 1910. https://doi.org/10.1080/03004430.2020.1803852

Elmer, S., Meyer, M., & Jäncke, L. (2012). Neurofunctional and behavioral correlates of phonetic and temporal categorization in musically trained and untrained subjects. *Cerebral Cortex (New York, N.Y.: 1991)*, *22*(3), 650–658. https://doi.org/10.1093/cercor/bhr142

Everest, F.A. & Pohlmann, K. (2009). *Master Handbook of Acoustics*, McGraw Hill Professional.

Fant, G. (1960). *Acoustic theory of speech production,* The Hague, The Netherlands, Mouton

Farnetani, E., Recasens, D. (1993). Anticipatory consonant-to-vowel coarticulation in the production of VCV sequences in Italian. *Lang. Speech*, 36, 279–302.

Farnetani, E. & Recasens, D.F. (2010). Coarticulation and Connected Speech Processes. *The Handbook of Phonetic Sciences*, 2nd Edition, 316 – 352. https://doi.org/10.1002/9781444317251.ch9

Feng, Y., Hao, G.J., Xue, S.A., Max, L. (2011) Detecting anticipatory effects in speech articulation by means of spectral coefficient analyses. *Speech Communication*, 53(6), 842-854. https://doi.org10.1016/j.specom.2011.02.003

Flege, J., Bohn, O.-S., & Jang, S. (1997). Effects of Experience on Non-Native Speakers' Production and Perception of English Vowels. *Journal of Phonetics, 25*, 437–470. https://doi.org/10.1006/jpho.1997.0052

Fogerty, D. & Humes, L.E. (2010). Perceptual contributions to monosyllabic word intelligibility: segmental, lexical, and noise replacement factors. *J Acoust Soc Am* 128, 3114–3125.

François, C. & Schön D. (2011). Musical expertise boosts implicit learning of both musical and linguistic structures. *Cereb Cortex*, 21(10), 2357- 2365. https://doi.org/10.1093/cercor/bhr022

Fritz, C., Curtin, J., Poitevineau, J., Borsarello, H., Wollman, I., Tao, F.C. & Ghasarossian, T. (2014). Soloist evaluations of six Old Italian and six new violins. *Proc Natl Acad Sci U S A*,111(20), 7224-7229. https://doi.org/10.1073/pnas.1323367111

Fujioka, T., Trainor, L.J., Ross, B., Kakigi, R. & Pantev, C. (2004). Musical training enhances automatic encoding of melodic contour and interval structure. *J Cogn Neurosci*, 16(6), 1010 - 1021. https://doi.org/10.1162/0898929041502706

Fujioka, T., Ross, B., Kakigi, R., Pantev, C., & Trainor, L. J. (2006). One year of musical training affects development of auditory cortical-evoked fields in young children. *Brain: A Journal of Neurology*, *129*(Pt 10), 2593–2608. https://doi.org/10.1093/brain/awl247

Gaser, C. & Schlaug, G. (2003). Brain structures differ between musicians and non-musicians. *J. Neurosci,* 23, 9240–9245. https://doi.org/10.1523/JNEUROSCI.23-27-09240.2003

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*, 613–656. https://doi.org/10.1080/016909697386646

Gelfand, S.A. (2016). *Hearing: An Introduction to Psychological and Physiological Acoustics*, CRC Press.

Gendrot, C. & Adda-Decker, M. (2005). Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. *Interspeech*, Lisbon, Portugal, 2453-2456.

Georgeton, L., Paillereau, N., Landron, S., Gao, J., & Kamiyama, T. (2012). Analyse formantique des voyelles orales du français en contexte isolé: À la recherche d'une référence pour les apprenants de FLE. 9. *Conférence conjointe JEP-TALN-RECITAL,* 142 -152.

Gordon, E. (1989). *Manual for the advanced measures of music education*, Chicago (IL) G.I.A. Publications, Inc.

Gordon, R. L., Magne, C. L., & Large, E. W. (2011). EEG Correlates of Song Prosody: A New Look at the Relationship between Linguistic and Musical Rhythm. *Frontiers in Psychology*, *2*, 352. https://doi.org/10.3389/fpsyg.2011.00352

Gordon, R. L., Fehd, H. M., & McCandliss, B. D. (2015). Does Music Training Enhance Literacy Skills? A Meta-Analysis. *Frontiers in Psychology*, *6*. https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01777

Gordon, R. & Magne, C. (2017). Music and Cognitive Abilities (Music in the Brain). *The Routledge Companion to Music Cognition*. New York: Routledge.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "l" and "r." *Neuropsychologia, 9*(3), 317–323. https://doi.org/10.1016/0028-3932(71)90027-3

Gottfried, T. L., and Riester, D. (2000). Relation of pitch glide perception and Mandarin tone identification. *J. Acoust. Soc. Am*, 108, 2604.

Gottfried, T. L., Staby, A. M., & Ziemer, C. J. (2004). Musical experience and Mandarin tone discrimination and imitation. *The Journal of the Acoustical Society of America*, *115*(5), 2545–2545. https://doi.org/10.1121/1.4783674

Gotzen, A. D., Bernadini, N. & Arfib, D. (2000). Traditional Implementations of a Phase Vocoder: The Tricks of the Trade. Proceedings of the *COST G-6 Conference on Digital Audio Effects (DAFX-00)*, Verona, Italy, December 7-9.

Gracyk, T. (2003). Does Everyone Have a Musical Indentity? Reflections on Musical Identities. *Action, Criticism & Theory for Music Education Electronic Article*.

Hallam, S. (2010). 21st century conceptions of musical ability. *Psychology of Music*, *38*(3), 308-330. https://doi.org/10.1177/0305735609351922

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2019). Mandarin Tone Identification by Tone-Naïve Musicians and Non-musicians in Auditory-Visual and Auditory-Only Conditions. *Frontiers in Communication*, *4*. https://doi.org/10.3389/fcomm.2019.00070

Hargreaves, D.J., Miell, D. & MacDonald, R.A. (2002). What are musical identities, and why are they important. *Musical Indentities*.

Hargreaves, D.J., MacDonald, R. & Miell, D. (2011). Musical identities mediate musical development. *Oxford Handbook of Music Education*.

Haynes, B. & Cooke, P. (2001). Pitch. *Grove Music Online*. https://doi.org/10.1093/gmo/9781561592630.article.40883

Healy, A. F., & Cutting, J. E. (1976). Units of speech perception: Phoneme and syllable. *Journal of Verbal Learning and Verbal Behavior*, *15*(1), 73–83. https://doi.org/10.1016/S0022-5371(76)90008-6

Herholz, S.C., Lappe, C., Knief, A. & Pantev, C. (2009). Imagery Mismatch Negativity in Musicians. *Annals of the New York Academy of Sciences*, 1169 (1), 173 – 177. https://doi.org/10.1111/j.1749-6632.2009.04782.x

Herholz, S. C., & Zatorre, R. J. (2012). Musical training as a framework for brain plasticity: Behavior, function, and structure. *Neuron*, *76*(3), 486–502. https://doi.org/10.1016/j.neuron.2012.10.011

Hervais-Adelman, A., Davis, M., Johnsrude, I. & Carlyon, R. (2008). Perceptual learning of noise vocoded words: effects of feedback and lexicality. *J Exp Psychol Hum Percept Perform*, 34(2), 460-474.

Hewlet, N. & Beck, J. (2010). An Introduction to the Science of Phonetics. *Taylor & Francis Group*.

Hickok, G. (2014). The architecture of speech production and the role of the phoneme in speech processing. *Language and Cognitive Processes*, *29*(1), 2–20. https://doi.org/10.1080/01690965.2013.834370

Hillenbrand, J.M., Clark, M.J. & Houde, R.A. (2000). Some effects of duration on vowel recognition. *J Acoust Soc Am*, 108(6), 3013-22. https://doi.org/10.1121/1.1323463

Ho, Y.-C., Cheung, M.-C., & Chan, A. S. (2003). Music training improves verbal but not visual memory: Cross-sectional and longitudinal explorations in children. *Neuropsychology*, *17*(3), 439–450. https://doi.org/10.1037/0894-4105.17.3.439

House, A. S. & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *J. Acoust. Soc. Am*. 25, 105–113. https://doi.org/10.1121/1.1906982

Hutka, S., Bidelman, G. M., & Moreno, S. (2015). Pitch expertise is not created equal: Cross-domain effects of musicianship and tone language experience on neural and behavioural

discrimination of speech and music. *Neuropsychologia*, *71*, 52–63. https://doi.org/10.1016/j.neuropsychologia.2015.03.019

Informations Générales 2022/2023, (2022). Retrieved November 3, 2022, from https://www.conservatoire-nice.org/docs/22-23/Preinscriptions2022-2023.pdf

Intartaglia, B., White-Schwoch, T., Kraus, N., & Schön, D. (2017). Music training enhances the automatic neural processing of foreign speech sounds. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-12575-1

Jakobson, L.S., Lewycky, S. T., Kilgour, A. & Stoesz, B. (2008). Memory for verbal and Visual Material in Highly Trained Musicians. *Music Perception: An Interdisciplinary Journal*, 26 (1), 41-55. https://doi.org/10.1525/mp.2008.26.1.41

Jones, J. L., Lucker, J., Zalewski, C., Brewer, C., & Drayna, D. (2009). Phonological processing in adults with deficits in musical pitch recognition. *Journal of Communication Disorders*, *42*(3), 226–234. https://doi.org/10.1016/j.jcomdis.2009.01.001

Kangatharan, J., Uther, M. & Gobet, F. (2022). The effect of hyperarticulation on speech comprehension under adverse listeing conditions, *Psychological Research*, 86, 1535 – 1546.

Kazanina, N., Bowers, J. S., & Idsardi, W. (2018). Phonemes: Lexical access and beyond. *Psychonomic Bulletin & Review*, *25*(2), 560–585. https://doi.org/10.3758/s13423-017-1362-0

Kleber, B., Veit, R., Birbaumer, N., Gruzelier, J. & Lotze, M. (2009). The Brain of Opera Singers : Experience-Dependent Changes in Functional Activation. *Cerebrebral Cortex*, 20(5), 1144-1152. https://doi.org/10.1093/cercor/bhp177

Koelsch, S., Schröger, E., & Tervaniemi, M. (1999). Superior pre-attentive auditory processing in musicians. *Neuroreport*, *10*(6), 1309–1313. https://doi.org/10.1097/00001756-199904260-00029

Koelsch, S. (2011). Towards a neural basis of music perception – a review and updated model. *Frontiers in Psychology*, 2 (110), 1-20, https://doi.org/10.3389/fpsyg.2011.00110

Kraus, N. & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nat Rev Neurosci* 11, 599–605. https://doi.org/10.1038/nrn2882

Krishnan, A., Bidelman, G.M., Smalt, C.J., Ananthakrishnan, S. & Gandour, J.T. (2012). Relationship between brainstem, cortical and behavioral measures relevant to pitch salience in humans. *Neuropsychologia*, 50(12), 2849-2859. https://doi.org/10.1016/j.neuropsychologia.2012.08.013

Kuang, J., & Liberman, M. (2018). Integrating voice quality cues in the pitch perception of speech and non-speech utterances. *Frontiers in Psychology, 9,* 2147. https://doi.org/10.3389/fpsyg.2018.02147

Kubik, S. (2016). *Inscriptions aux conservatoires: Comment choisir son instrument ?* France Musique. https://www.radiofrance.fr/francemusique/inscriptions-aux-conservatoires-comment-choisir-son-instrument-4104428

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000. https://doi.org/10.1098/rstb.2007.2154

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29,* 98–104. https://doi.org/10.1121/1.1908694

Ladefoged, P. (1993) A Course in Phonetics, *Harcourt Brace Jovanovich College Publishers.*

Lappe, C., Herholz, S. C., Trainor, L. J., & Pantev, C. (2008). Cortical Plasticity Induced by Short-Term Unimodal and Multimodal Musical Training. *Journal of Neuroscience*, *28*(39), 9632–9639. https://doi.org/10.1523/JNEUROSCI.2254-08.2008

Lappe, C., Trainor, L. J., Herholz, S. C., & Pantev, C. (2011). Cortical Plasticity Induced by Short-Term Multimodal Musical Rhythm Training. *PLOS ONE*, *6*(6), e21493. https://doi.org/10.1371/journal.pone.0021493

Latinus, M. & Belin, P. (2011). Human voice perception. *Curr Biol,* 21(4), 143-145. https://doi.org/10.1016/j.cub.2010.12.033

Lee, C.Y., & Hung, T.H. (2008). Identification of Mandarin tones by English-speaking musicians and non-musicians. *The Journal of the Acoustical Society of America*, *124*(5), 3235–3248. https://doi.org/10.1121/1.2990713

Lee, C.Y., & Lee, Y.F. (2010). Perception of musical pitch and lexical tones by Mandarin-speaking musicians. *The Journal of the Acoustical Society of America*, *127*(1), 481–490. https://doi.org/10.1121/1.3266683

Leite, R.B., Mota-Rolim, S.A., Queiroz, C.M. (2016). Music Proficiency and Quantification of Absolute Pitch: A Large-Scale Study among Brazilian Musicians. *Front Neurosc*i, 10 (447). https://doi.org/10.3389/fnins.2016.00447

Lerdahl, F. & Jackendoff, R. (1983). *A Generative Theory of Tonal Music.* The MIT Press.

Leroy, C. (1970). Étude de phonétique comparative de la langue turque sifflée et parlée. *Revue de Phonétique Appliquée*, 14/15, 119-161.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *The Behavioral and Brain Sciences*, *22*(1), 1–38. https://doi.org/10.1017/s0140525x99001776

Liang, C., Earl, B., Thompson, I., Whitaker, K., Cahn, S., Xiang, J., Fu, Q.J. & Zhang, F. (2016). Musicians Are Better than Non-musicians in Frequency Change Detection: Behavioral and Electrophysiological Evidence. *Front Neurosci*, 25(10), 464. https://doi.org/10.3389/fnins.2016.00464

Liberman, A. M., Delattre, P.C. & Cooper, F.S. (1952). The role of selected stimulus variables in the perception of the unvoiced stop consonants. *Am J Psychol*, 65, 497–516.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied, 68*(8), 1–13. https://doi.org/10.1037/h0093673

Liberman, A. M., F. S. Cooper, D. P. Shankweiler & M. Studdert-Kennedy. (1967). Perception of the speech code. *Psychological Review*, 74(6). 431–461.

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

Lipski, J. (n.d.) *THE SPANISH OF THE CANARY ISLANDS*. Retrieved September 26, 2022, from http://www.personal.psu.edu/jml34/Canary.htm

Listes d'attente: Une fatalité ? (2015, December 14). *Conservatoires de France*. https://conservatoires-de-france.com/listes-dattente-fatalite/

Logan, C. (2014). *Action and Instrument Specificity in Musicians*. Thesis. University of Connecticut.

Lotto, A.J., Sato, M., & Diehl, R.L. (2004). Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. In J. Slifka, S. Manuel, & M. Matthies (Eds.), *From sound to sense: 50+ years of discoveries in speech communication*, 181 – 186.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, *19*(1), 1–36.

Magnussen, J., Mirman, D. & Harris, H. (2012). Computational models of spoken word recognition. In M. J. Spivey, K. McRae, & M. F. Joanisse (Eds.), *The Cambridge handbook of psycholinguistics*, 76–103. https://doi.org/10.1017/CBO9781139029377.008

Molina Mejia, J. (2007) *Diagnostique et Correction des Erreurs de Prononciation en FLE des apprenants Hispanophones*, Mémoire M1, direction de Dominique Abry, Grenoble : Université de Stendhal.

Margulis, E.H., Mlsna, L.M., Uppunda, A.K., Parrish, T.B. & Wong, P.C. (2009). Selective neurophysiologic responses to music in instrumentalists with different listening biographies. *Hum Brain Mapp*, 30(1), 267-75. https://doi.org/10.1002/hbm.20503

Marie, C., Delogu, F., Lampis, G., Belardinelli, M. O., & Besson, M. (2011). Influence of musical expertise on segmental and tonal processing in Mandarin Chinese. *Journal of Cognitive Neuroscience*, *23*(10), 2701–2715. https://doi.org/10.1162/jocn.2010.21585

Marques, C., Moreno, S., Castro, S. L., & Besson, M. (2007). Musicians detect pitch violation in a foreign language better than nonmusicians: Behavioral and electrophysiological evidence. *Journal of Cognitive Neuroscience*, *19*(9), 1453–1463. https://doi.org/10.1162/jocn.2007.19.9.1453

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63. https://doi.org/10.1016/0010-0285(78)90018-X

Martínez-Montes, E., Hernández-Pérez, H., Chobert, J., Morgado-Rodríguez, L., Suárez-Murias, C., Valdés-Sosa, P. A., & Besson, M. (2013). Musical expertise and foreign speech perception. *Frontiers in Systems Neuroscience*, *7*, 84. https://doi.org/10.3389/fnsys.2013.00084

Massapallo, M., Polka, L. & Ménard, L. (2017). A universal bias in adult vowel perception – By ear or by eye. *Cognition*, 358-370. https://doi.org/10.1016/j.cognition.2017.06.001

Massaro, D. W. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, *79*, 124–145. https://doi.org/10.1037/h0032264

Massaro, D. W. (1974). Perceptual Units in Speech Recognition. *Journal of Experimental Psychology*, 102(2), 349-353.

Massaro, D.W. (1975). Understanding Language: An Information Processing Analysis of Speech Perception, Reading and Psycholinguistics. *New York: Academic Press*. https://doi.org/10.1016/B978-0-12-478350-8.50006-4

Mateos-Aparicio, P. & Rodríguez-Moreno, A. (2019). The Impact of Studying Brain Plasticity. *Frontiers in Cellular Neuroscience*, *13*. https://doi.org/10.3389/fncel.2019.00066

Mattys, S. L., & Melhorn, J. F. (2005). How do Syllables Contribute to the Perception of Spoken English? Insight from the Migration Paradigm. *Language and Speech, 48(2)*, 223-252. https://doi.org/10.1177/00238309050480020501

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McPherson, L., & Winter, Y. (2022). Editorial: Surrogate Languages and the Grammar of Language-Based Music. *Frontiers in Communication*, *7*. https://doi.org/10.3389/fcomm.2022.838286

McQueen, J. M. (2005). Speech perception. *The Handbook of Cognition,* 255–275. https://doi.org/10.4135/9781848608177.n11

Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, *20*(3), 298–305. https://doi.org/10.1016/S0022-5371(81)90450-3

Mehler, J., Sebastian-Gallés, N., Altmann, G., Dupoux, E., Christophe, A. & Pallier, C. (1993). "Understanding compressed sentences: the role of rhythm and meaning". In A. M. G. Paula

Tallal, Rodolfo R. Llinas, Curt von Euler (Ed.), *Temporal information processing in the nervous system: Special reference to dyslexia and dysphasia. Annals of the New York Academy of Sciences*, Vol. 682, 272-282.

Meunier, C. (2007). Phonétique acoustique. Auzou P. *Les dysarthries*, Solal, 164-173.

Meyer, L. B. (1973). Explaining music: Essays and explorations. *Berkeley: University of California Press*.

Meyer, J. (2005). *Description typologique et intelligibilité des langues sifflées, approche linguistique et bioacoustique.* Thesis. University Lyon 2.

Meyer, J. (2008). Acoustic Strategy and Typology of Whistled Languages; Phonetic Comparison and Perceptual Cues of Whistled Vowels. *Journal of the International Phonetic Association*. Cambridge University Press, 38 (1), 69-94.

Meyer, J., Dentel, L., Meunier, F., (2013). "Speech Recognition in Natural Background Noise". *PLoS ONE*, 8(11): e79279.

Meyer, J. (2015). *Whistled Languages, A Worldwide Inquiry on Human Whistled Speech*. Springer.

Meyer, J., Dentel, L., & Meunier, F. (2017). Categorization of Natural Whistled Vowels by Naïve Listeners of Different Language Background. *Frontiers in Psychology*, *8*. https://www.frontiersin.org/articles/10.3389/fpsyg.2017.00025

Meyer, J., & Díaz, D. (2017). Geolingüística de los lenguajes silbados del mundo, con un enfoque en el español silbado. *Géolinguistique*, *17*, Article 17. https://doi.org/10.4000/geolinguistique.373

Meyer, J., Meunier, F., Dentel, L., Do Carmo Blanco, N. & Sèbe, F. (2018). Loud and Shouted Speech Perception at Variabile Distances in a Forest. *Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association*, 2285-2289. https://doi.org/10.21437/interspeech.2018-2089

Meyer, J., Dentel, L., Gerber, S., Ridouane, R. (2019) A Perceptual Study of CV Syllables in Both Spoken and Whistled Speech: A Tashlhiyt Berber Perspective. *Proc. Interspeech 2019*, 2295-2299. https://doi.org/10.21437/Interspeech.2019-2251

Meyer, J. (2021a) Environmental and linguistic typology of whistled languages. *Annual Review of Linguistics*, 7, 493-510.

Meyer, J., Magnasco, M. O., & Reiss, D. (2021b). The Relevance of Human Whistled Languages for the Analysis and Decoding of Dolphin Communication. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.689501

Miendlarzewska, E. A., & Trost, W. J. (2014). How musical training affects cognitive development: Rhythm, reward and other modulating variables. *Frontiers in Neuroscience*, *7*, 279. https://doi.org/10.3389/fnins.2013.00279

Mills, M. (2010). Being a Musician: Musical Identity and the Adolescent Singer. *Bulletin of the Council for Research in Music Education*, 186, 43-54.

Milovanov, R., Pietilä, P., Tervaniemi, M., & Esquef, P. A. A. (2010). Foreign language pronunciation skills and musical aptitude: A study of Finnish adults with higher education. *Learning and Individual Differences*, *20*(1), 56–60. https://doi.org/10.1016/j.lindif.2009.11.003

Mirenda, P. & Beukelman, D.R. (1987). A comparison of speech synthesis intelligibility with listeners from three age groups. *Augmentative and Alternative Communication,* 3, 120–128.

Mitterer, H., & Müsseler, J. (2013). Regional accent variation in the shadowing task: Evidence for a loose perception–action coupling in speech. *Attention, Perception, & Psychophysics*, *75*(3), 557–575. https://doi.org/10.3758/s13414-012-0407-8

Mitterer, H., Reinisch, E., & McQueen, J. M. (2018). Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language*, *98*, 77–92. https://doi.org/10.1016/j.jml.2017.09.005

Miyawaki, K., Jenkins, J.J., Strange, W. Liberman, A.M., Verbrugge, R. & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics* 18, 331–340. https://doi.org/10.3758/BF03211209

Moles, A. (1970). Etude sociolinguistique de la langue sifflée de Kusköy. *Revue de Phonétique Appliquée*, 14/15, 78-118.

Montandon, F. (2018). Les représentations sociales de l'instrument de musique, objet sonore et espace de création identitaire. *Spécificités*, *11*(1), 48–61. https://doi.org/10.3917/spec.011.0048

Morais, J., Castro, S.L., Scliar-Cabral, L., Kolinsky, R. & Content, A. (1987). The effects of literacy on the recognition of dichotic words, *The Quarterly Journal of Experimental Psychology Section A*, 39(3), 451-465, https://doi.org/10.1080/14640748708401798

Moreno, S., Marques, C., Santos, A., Santos, M., Castro, S. L., & Besson, M. (2009). Musical training influences linguistic abilities in 8-year-old children: More evidence for brain plasticity. *Cerebral Cortex (New York, N.Y.: 1991)*, *19*(3), 712–723. https://doi.org/10.1093/cercor/bhn120

Moskowitz, B. A. (1975). The acquisition of Fricatives: a study in phonetics and phonology, *J. Phonetics*, 3(3), 141–150.

New, B. & Pallier, C. (2023). Lexique. http://www.lexique.org/

Newport, E. & Aslin, R. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology* 48, 127-162.

Nguyen, N. (2001). Rôle de la coarticulation dans la reconnaissance des mots. *L'année psychologique*, 101 (1), 125-154. https://doi.org/10.3406/psy.2001.29719

Nie, P., Wang, C., Rong, G., Du, B., Lu, J., Li, S., Putkinen, V., Tao, S. & Tervaniemi, M. (2022) Effects of Music Training on the Auditory Working Memory of Chinese-Speaking School-Aged Children: A Longitudinal Intervention Study. *Front. Psychol,*12, https://doi.org/10.3389/fpsyg.2021.770425

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189–234. https://doi.org/10.1016/0010-0277(94)90043-4

Öhman, S.E.G (1966). Coarticulation in VCV Utterances: Spectrographic Measurements. Journal of Acoustic Society of America 39, 151-168. https://doi.org/10.1121/1.1909864

Olszewska, A.M., Gaca, M., Herman, A.M., Jednoróg, K. & Marchewka, A. (2021). How Musical Training Shapes the Adult Brain: Predispositions and Neuroplasticity. *Front. Neurosci,* 15:630829. https://doi.org/10.3389/fnins.2021.630829

Ong, J. H., Burnham, D., Escudero, P., & Stevens, C. J. (2017). Effect of Linguistic and Musical Experience on Distributional Learning of Nonnative Lexical Tones. *Journal of Speech, Language, and Hearing Research*, *60*(10), 2769–2780. https://doi.org/10.1044/2016_JSLHR-S-16-0080

Ott, C. G. M., Langer, N., Oechslin, M. S., Meyer, M., & Jäncke, L. (2011). Processing of Voiced and Unvoiced Acoustic Stimuli in Musicians. *Frontiers in Psychology*, *2*, 195. https://doi.org/10.3389/fpsyg.2011.00195

Paillereau, N.M. (2015). *Perception et production des voyelles orales du francais par des futures enseignantes tchèques de Francais Langue Etrangère (FLE)*. Thèse. Université de la Sorbonne nouvelle - Paris III.

Paliwal, K. K., Ainsworth, W. A. & Lindsay, D. (1983), A study of two-formant models for vowel identification. *Speech Communication*, 2(4), 295-303. https://doi.org/10.1016/0167-6393(83)90046-8

Pantev, C., Oostenveld, R., Engelien, A., Ross, B., Roberts, L.E. & Hoke, M. (1998). Increased auditory cortical representation in musicians. *Nature*, 392(6678), 811-814. https://doi.org/10.1038/33918

Pantev, C., Roberts, L.E., Schulz, M., Engelien, A. & Ross, B. (2001). Timbre-specific enhancement of auditory cortical representations in musicians. *Neuroreport*, 12(1), 169-174. https://doi.org/10.1097/00001756-200101220-00041

Pantev, C., & Herholz, S. C. (2011). Plasticity of the human auditory cortex related to musical training. *Neuroscience and Biobehavioral Reviews*, *35*(10), 2140–2154. https://doi.org/10.1016/j.neubiorev.2011.06.010

Parbery-Clark, A., Skoe, E., & Kraus, N. (2009a). Musical experience limits the degradative effects of background noise on the neural processing of sound. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(45), 14100–14107. https://doi.org/10.1523/JNEUROSCI.3256-09.2009

Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009b). Musician enhancement for speech-in-noise. *Ear and Hearing*, *30*(6), 653–661. https://doi.org/10.1097/AUD.0b013e3181b412e9

Parbery-Clark, A., Tierney, A., Strait, D.L., Kraus, N. (2012). Musicians have fine-tuned neural distinction of speech syllables. *Neuroscience*,111-9. https://doi.org/10.1016/j.neuroscience.2012.05.042

Pégourdie, A. (2015). L'« instrumentalisation » des carrières musicales Division sociale du travail, inégalités d'accès à l'emploi et renversement de la hiérarchie musicale dans les conservatoires de musique. *Sociologie*, *6*(4), 321–338. https://doi.org/10.3917/socio.064.0321

Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience*, *6*(7). https://doi.org/10.1038/nn1083

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *The Journal of the Acoustical Society of America*, *116*(4 Pt 1), 2338–2344. https://doi.org/10.1121/1.1787524

Pierrehumbert, J. B. (2003). Phonetic Diversity, Statistical Learning, and Acquisition of Phonology, *Language and Speech*, 2003, 46(2-3), 115 – 154.

Pisoni, D.B. (1979). On the perception of speech sounds as biologically significant signals. *Brain Behav Evol*, 16(5-6), 330-50. https://doi.org/10.1159/000121875

Polka, L., Molnar, M., Zhao, T.C. & Masapollo, M. (2021). Neurophysiological Correlates of Asymmetries in Vowel Perception: An English-French Cross-Linguistic Event-Related Potential Study, *Front. Hum. Neurosci.* https://doi.org/10.3389/fnhum.2021.607148

Proverbio, A. M., & Orlandi, A. (2016). Instrument-specific effects of musical expertise on audiovisual processing (Clarinet vs. Violin). *Music Perception, 33*(4), 446–456. https://doi.org/10.1525/mp.2016.33.4.446

Quedenfeldt, H. M. (1887). Pfeifsprache auf der Insel Gomera. *Zeitschrift für Ethnologie*, 19, 731-741.

Rammell, C. S., Cheng, H., Pisoni, D. B., & Newman, S. D. (2019). L2 speech perception in noise: An fMRI study of advanced Spanish learners. *Brain Research*, *1720*, 146316. https://doi.org/10.1016/j.brainres.2019.146316

Recasens, D. (1987). An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences. *Journal of Phonetics*, 15, 299–312.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carell, T. D. (1981). Speech perception without traditional speech cues. Science 212, 947–950. https://doi.org/10.1126/science.7233191

Reymore, L. (2021). Characterizing prototypical musical instrument timbres with Timbre Trait Profiles. *Musicae Scientiae*, 1 – 27. https://doi.org/10.1177/10298649211001523

Rialland, A. (2005). Phonological and phonetic aspects of whistled languages. *Phonology*, Cambridge University Press (CUP), 22 (2), 237-271.

Ridouane, R., Turco, G. & Meyer, J. (2018). Length Contrast and Covarying Features: Whistled Speech as a Case Study. *Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association, Hyperabad, India*, 1843-1847. https://doi.org/10.21437/interspeech.2018-1060

Rigoulot, S., Pell, M. D., and Armony, J. L. (2015). Time course of the influence of musical expertise on the processing of vocal and musical sounds. *Neuroscience* 290, 175–184. https://doi.org/10.1016/j.neuroscience.2015.01.033

Rodriguez-Gomez, D.A. & Talero-Gutiérrez, C. (2022) Effects of music training in executive function performance in children: A systematic review. *Front. Psychol,* 13. https://doi.org/10.3389/fpsyg.2022.968144

Rognoni, G. R. (2008). The Development Of Organology as a Discipline Through Early Museum Catalogues. *English Translation by Julia Weiss of La Definizione Dell'organologia Come Disciplina..., "Annali, Nuova Serie", IX (2008), Florence : Università Degli Studi Di Firenze – Dipartimento Di Storia Delle Arti e Dello Spettacolo – Titivillus*, 155-171.

Román-Caballero, R., Vadillo, M. A., Trainor, L. J., & Lupiáñez, J. (2022). Please don't stop the music: A meta-analysis of the cognitive and academic benefits of instrumental musical training in childhood and adolescence. *Educational Research Review*, *35*, 100436. https://doi.org/10.1016/j.edurev.2022.100436

Ross, D. A., Gore, J. C., & Marks, L. E. (2005). Absolute pitch: Music and beyond. *Epilepsy & Behavior: E&B*, *7*(4), 578–601. https://doi.org/10.1016/j.yebeh.2005.05.019

Ruggles, D. R., Freyman, R. L., & Oxenham, A. J. (2014). Influence of Musical Training on Understanding Voiced and Whispered Speech in Noise. *PLOS ONE*, *9*(1), e86980. https://doi.org/10.1371/journal.pone.0086980

Sadakata, M., & Sekiyama, K. (2011). Enhanced perception of various linguistic features by musicians: A cross-linguistic study. *Acta Psychologica*, *138*(1), 1–10. https://doi.org/10.1016/j.actpsy.2011.03.007

Sala, G., Aksayli, N. D., Tatlidil, K. S., Tatsumi, T., Gondo, Y., & Gobet, F. (2019). Near and Far Transfer in Cognitive Training: A Second-Order Meta-Analysis. *Collabra: Psychology*, *5*(1), 18. https://doi.org/10.1525/collabra.203

Sala, G., & Gobet, F. (2017). Does Far Transfer Exist? Negative Evidence From Chess, Music, and Working Memory Training. *Current Directions in Psychological Science*, *26*(6), 515–520. https://doi.org/10.1177/0963721417712760

Sala, G., & Gobet, F. (2020). Cognitive and academic benefits of music training with children: A multilevel meta-analysis. *Memory & Cognition*, *48*(8), 1429–1441. https://doi.org/10.3758/s13421-020-01060-2

Sammler, D. & Elmer, S. (2020) Advances in the Neurocognition of Music and Language. *Brain Sci.* 10(8): 509. https://doi.org/10.3390/brainsci10080509

Scarborough, R. & Zellou, G. (2013). Clarity in communication: "clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *J Acoust Soc Am*. Nov. 134(5), 3793-807. https://doi.org/10.1121/1.4824120

Schellenberg, E. G., & Weiss, M. W. (2013). Music and cognitive abilities. In D. Deutsch (Ed.), *The psychology of music* (3rd ed.), 499–550. Amsterdam: Elsevier.

Schlaug, G., Norton, A., Overy, K. & Winner, E. (2005). Effects of music training on the child's brain and cognitive development. *Ann N Y Acad Sci*, 1060, 219-230. https://doi.org/10.1196/annals.1360.015

Schön, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology*, *41*(3), 341–349. https://doi.org/10.1111/1469-8986.00172.x

Schwartz, J.L & Escudier, P. (1989). A strong evidence for the existence of a large scale integrated spectral representation in vowel perception, *Speech Commun*. 8, pp. 235–259.

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1998). The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics*, 25, 3, 255-286. https://doi.org/10.1006/jpho.1997.0043

Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *25*(5), 336–354. https://doi.org/10.1016/j.jneuroling.2009.12.004

Semal, C., Demany, L., Ueda, K., & Hallé, P. A. (1996). Speech versus nonspeech in pitch memory. *The Journal of the Acoustical Society of America*, *100*(2 Pt 1), 1132–1140. https://doi.org/10.1121/1.416298

Sergeant, D.C. & Himonides, E. (2019) Orchestrated sex: The representation of male and female musicians in world-class symphony orchestras. *Frontiers in Psychology*, 10, 1760 https://doi.org/10.3389/fpsyg.2019.01760

Shadle, C. H. (1983). Experiments on the Acoustics of Whistling. *Physics Teacher*, *21*(3), 148–154.

Shahin, A. J., Roberts, L. E., Chau, W., Trainor, L. J., & Miller, L. M. (2008). Music training leads to the development of timbre-specific gamma band activity. *NeuroImage*, *41*(1), 113–122. https://doi.org/10.1016/j.neuroimage.2008.01.067

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J. & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234), 303-304. https://doi.org/10.1126/science.270.5234.303

Siedenburg, K. & McAdams, S. (2017). Four Distinctions for the Auditory "Wastebasket" of Timbre. *Front Psychol,* 8, 1747. https://doi.org/10.3389/fpsyg.2017.01747

Siedenburg, K., Saitis, C. & McAdams, S. (2019). The Present, Past, and Future of Timbre Research. *Timbre: Acoustics, Perception, and Cognition*. Springer Handbook of Auditory Research, vol 69, 1-19. https://doi.org/10.1007/978-3-030-14832-4_1

Skubic, D., Gaberc, B., & Jerman, J. (2021). Supportive Development of Phonological Awareness Through Musical Activities According to Edgar Willems. *SAGE Open*, *11*(2). https://doi.org/10.1177/21582440211021832

Slevc, L. R., & Miyake, A. (2006). Individual Differences in Second-Language Proficiency: Does Musical Ability Matter? *Psychological Science*, *17*(8), 675–681. https://doi.org/10.1111/j.1467-9280.2006.01765.x

Smit, E., Rathcke, T. & Keller, P. (2023). Tuning the Musical Mind: Next Steps in Solving the Puzzle of the Cognitive Transfer of Musical Training to Language and Back. *Music & Science* (6). https://doi.org/10.1177/20592043231175251

Solé, M. J., & Ohala, J. J. (2010). What is and what is not under the control of the speaker. Intrinsic vowel duration. *Papers in Laboratory Phonology 10*, eds C. Fougeron, B. Kühnert, M. D'Imperio, and N. Vallée (Berlin: de Gruyter).

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, 17, 3-45. https://doi.org/10.1016/S0095-4470(19)31520-7

Stevens, K. N. (1998). Acoustic phonetics. *MIT Press,* Cambridge, MA.

Stoll, G. (1984). Pitch of vowels: Experimental and theoretical investigation of its dependence on vowel quality. *Speech Communication, 3*(2), 137–150. https://doi.org/10.1016/0167-6393(84)90035-9

Strait, D. L., & Kraus, N. (2011). Can you hear me now? Musical training shapes functional brain networks for selective auditory attention and hearing speech in noise. *Frontiers in Psychology*, *2*, 113. https://doi.org/10.3389/fpsyg.2011.00113

Strange, W. (1989). Evolving theories of vowel perception. *Journal of Acoustic Societies of America*. 2081 – 2087.

Sundberg, J. (2003). Research on the singing voice in retrospect. *Department of Speech, Music, and Hearing Quarterly Progress and Status Report,* 45(1), 11–22.

Swaminathan, S., & Schellenberg, E. G. (2017). Musical competence and phoneme perception in a foreign language. *Psychonomic Bulletin & Review*, *24*(6), 1929–1934. https://doi.org/10.3758/s13423-017-1244-5

Syrdal, A.K. & Gopal, H.S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *J Acoust Soc Am*. 79(4),1086-100. https://doi.org/10.1121/1.393381

Tableau des cycles d'études en conservatoire. (n.d.). Retrieved November 28, 2022, from https://metiers.philharmoniedeparis.fr/tableau-cycles-etudes-conservatoire.aspx

Tang, W., Xiong, W., Zhang, Y.-X., Dong, Q., & Nan, Y. (2016). Musical experience facilitates lexical tone processing among Mandarin speakers: Behavioral and neural evidence. *Neuropsychologia*, *91*, 247–253. https://doi.org/10.1016/j.neuropsychologia.2016.08.003

van Tassell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P.(1987). Speech waveform envelope cues for consonant recognition. *J. Acoust. Soc. Am*. 82, 1152–1161. https://doi.org/10.1121/1.395251

Tervaniemi, M., Just, V., Koelsch, S., Widmann, A., & Schröger, E. (2005). Pitch discrimination accuracy in musicians vs nonmusicians: An event-related potential and behavioral study. *Experimental Brain Research*, *161*(1), 1–10. https://doi.org/10.1007/s00221-004-2044-5

Tervaniemi, M., Janhunen, L., Kruck, S., Putkinen, V., Huotilainen, M. (2016). Auditory Profiles of Classical, Jazz, and Rock Musicians: Genre-Specific Sensitivity to Musical Sound Features. *Front Psychol*, 6:1900. https://doi.org/10.3389/fpsyg.2015.01900

Tierney, A. T., Bergeson-Dana, T. R., & Pisoni, D. B. (2008). Effects of Early Musical Experience on Auditory Sequence Memory. *Empirical Musicology Review : EMR*, *3*(4), 178–186.

Toro, J.M., Nespor, M., Mehler & J., Bonatti, L.L. (2008). Finding Words and Rules in a Speech Stream: Functional Differences Between Vowels and Consonants, *Psychological Science*, 19 (2), 137-144. https://doi.org/10.1111/j.1467-9280.2008.02059

Tranchant, L. (2016). Des musiciens à bonne école Les pratiques éducatives des classes supérieures au prisme de l'apprentissage enfantin de la musique. *Sociologie*, *7*(1), 23–40. https://doi.org/10.3917/socio.071.0023

Tran Ngoc, A., Meyer, J. and Meunier, F. (2020a). Categorization of Whistled Consonants by Naïve French Speakers. *INTERSPEECH 2020,* 1600-1604. https://doi.org/10.21437/Interspeech.2020-2683

Tran Ngoc, A., Meyer, J. and Meunier, F. (2020b).Whistled Vowel Identification by French Speakers. *INTERSPEECH 2020*, 1605-1609. https://doi.org/10.21437/Interspeech.2020-2697

Tran Ngoc, A., Meunier, F. & Meyer, J. (2022a). Testing perceptual flexibility in speech through the categorization of whistled Spanish consonants by French speakers. *JASA Letters*. https://doi.org/10.1121/10.0013900

Tran Ngoc, A., Meyer, J. and Meunier, F. (2022b) Bénéfices de la pratique musicale sur la catégorisation de la parole sifflée : analyse des processus de transferts. *Proc. XXXIVe Journées d'Études sur la Parole -- JEP 2022*, 405-413. https://doi.org/10.21437/JEP.2022-43

Tran Ngoc, A., Meunier, F., Meyer, J. (2023a). The Effect of Whistled Vowels on Whistled Word Categorization for Naive Listeners. *Proc. INTERSPEECH 2023*, 3063-3067. https://doi.org/10.21437/Interspeech.2023-1967

Traube, C. (2015). La notation du timbre instrumental: Noter la cause ou l'effet dans le rapport geste-son. *Circuit*, *25*(1), 21–37. https://doi.org/10.7202/1029474ar

Trujillo, R. (1978). *El silbo gomero. Análisis lingüístico*. Interinsular Canaria.

Trujillo, C. R., (2006), *El silbo gomero. Nuevo estudio fonológico* (edición bilingüe español-inglés), Academia Canaria de la Lengua.Trujillo, C. R. (1978). *El Silbo Gomero: Analisis linguistico*. Santa Cruz de Tenerife: Andres Bello

Tuley, S. (2021). A Study of Language and Its Uses in Flute Performance and Pedagogy. Theses and Disserations – Music, 183. https://uknowledge.uky.edu/music_etds/183

Uther, M., Giannakopoulou, A. & Iverson, P. (2012). Hyperarticulation of vowels enhances phonetic change responses in both native and non-native speakers of English: Evidence from an auditory event-related potential study. *Brain Research*, 1470, 52–58.

Vaissière, J. (2020). *La Phonétique*. Presses Universitaires de France.

Van Zuijen, T. L., Sussman, E., Winkler, I., Näätänen, R., & Tervaniemi, M. (2004). Grouping of Sequential Sounds—An Event-Related Potential Study Comparing Musicians and Nonmusicians. *Journal of Cognitive Neuroscience*, *16*, 331–338. https://doi.org/10.1162/089892904322984607

Varnet, L., Meyer, J., Hoen, M. & Meunier, F. (2012). Phoneme resistance during speech-in-speech comprehension, *Proceedings of Interspeech Portland, USA*.

Varnet, L., Wang, T., Peter, C., Meunier, F., & Hoen, M. (2015). How musical expertise shapes speech perception: Evidence from auditory classification images. *Scientific Reports*, *5*(1), Article 1. https://doi.org/10.1038/srep14489

Verdis, A., & Sotiriou, C. (2018). The psychometric characteristics of the Advanced Measures of Music Audiation in a region with strong non-Western music tradition. *International Journal of Music Education*, *36*, 69–84. https://doi.org/10.1177/0255761417689925

Villacorta, V., Perkell, J. S. & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception, *PubMed*. https://pubmed.ncbi.nlm.nih.gov/17902866/

von Helmholtz H. (1885, 1954). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. New York, NY: Dover; trans. by Ellis A. J.

Vuust, P., Brattico, E., Seppänen, M., Näätänen, R., & Tervaniemi, M. (2012). The sound of music:

Wang, W. S.-Y. (1967). Phonological features of tone. *Int. J. Am. Ling.* 33, 93–105. https://doi/10.1086/464946

Wang, D. & Zheng, T.F. (2015). Transfer Learning for speech and Language Processing , *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity. *Memory & Cognition*, *38*(2), 163–175. https://doi.org/10.3758/MC.38.2.163

Wing, H. D. (1962). A Revision of the "Wing Musical Aptitude Test." *Journal of Research in Music Education*, *10*(1), 39–46. https://doi.org/10.2307/3343909

Zaar, J. & Dau, T. (2015). Sources of variability in consonant perception of normal hearing listeners. *The Journal of the Acoustical Society of America*, 138(3), 1253-1268. https://doi.org/10.1121/1.4928142

Zarate, M. J., Ritson, C. R., & Poeppel, D. (2012). Pitch-interval discrimination and musical expertise: Is the semitone a perceptual boundary? *The Journal of the Acoustical Society of America*, *132*(2), 984–993. https://doi.org/10.1121/1.4733535

Zhang J. D., Susino M., McPherson G. E. & Schubert E. (2018). The definition of a musician in music psychology: A literature review and the six-year rule. *Psychology of Music*, 48(3), 389–409. https://doi.org/10.1177/0305735618804038

Zhang, J.D. & Schubert, E. (2019). A single Item Measure for Identifying Musician and Nonmusician categories based on Measures of Musical Sophistication. *Music Perception* (2019) 36 (5): 457–467. https://doi.org/10.1525/mp.2019.36.5.457

Zhang, J. D., Susino, M., McPherson, G. E., & Schubert, E. (2020). The definition of a musician in music psychology: A literature review and the six-year rule. *Psychology of Music*, *48*(3), 389-409. https://doi.org/10.1177/0305735618804038

Zhao, T.C, Masapollo, M., Polka, L., Ménard, L., & Kuhl, P.K. (2019). Effects of formant proximity and stimulus prototypicality on the neural discrimination of vowels: Evidence from the auditory frequency-following response. *Brain Lang,* 77-83. https://doi.org/10.1016/j.bandl.2019.05.002

Annex

A.1 – Chapter 1

Table 1:

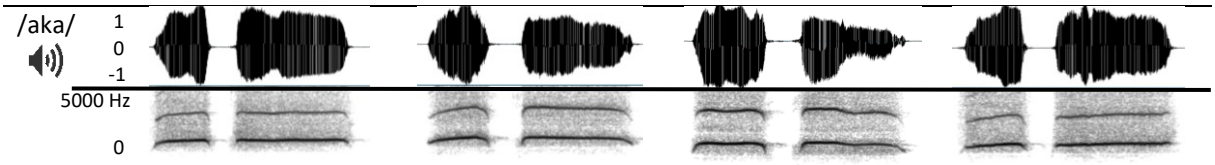
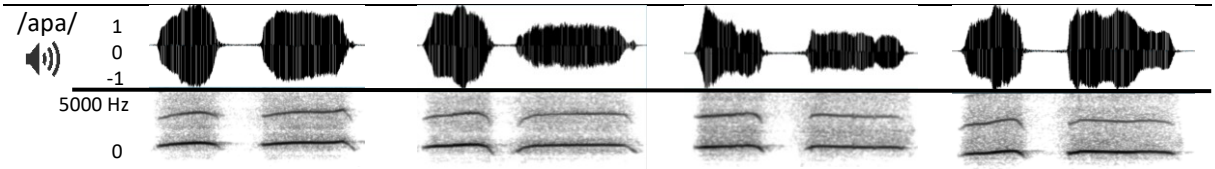
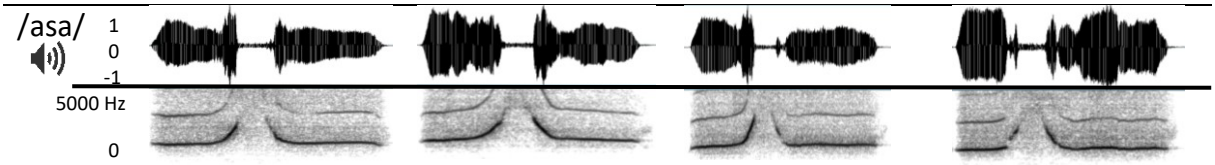
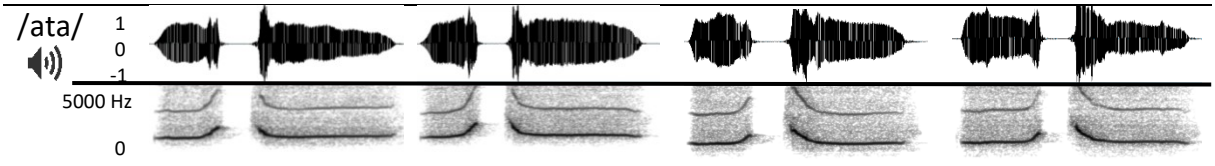
Qualifying musician participants

	Source		Yrs of training	Min age	Still playing	Instrument (Style)
Group 1 - Questionnaire	Ong, et al., 2017		2	-	-	-
	Sadakata & Sekiyama, 2011		5	-	yes	piano, flute, violin, clarinette, sax…
	Chan et al., 1998		6	12	-	-
	Varnet et al., 2015		7	13	-	-
	Williamson et al., 2010		8	-	-	(ability to read music)
	Bidelman & Krishnan, 2010		10	11	yes	trumpet, sax, piano, bass violin, tuba..
	Parbery-Clark et al., 2009a		10	7	-	-
	Dittinger et al., 2016		11	-	yes	piano, accordéon, violin, cello, guitar..
	François & Schön, 2011		12	-	yes	-
	Marie et al., 2011		*M* = 16	*M* = 7	-	sax, bass, violin, piano, organ, harp…
	Swaminathan & Schellenberg, 2017		*M* = 4.9	-	-	-
	Danielsen et al., 2022		pro	0	Yes	(jazz, trad, pop)
Group 2 - Diploma	Tang et al., 2016		6	8	yes	piano, flute (Classical), guzheng (Traditional)
	Liang et al., 2016		10	7	-	piano, guitar, sax, cello, trumpet, horn, bass (Western Classical)
	Ruggles et al., 2014		10	10	yes	x
	Martínez-Montes et al., 2013		12	7	yes	piano, violin, conducting, composition, singing, bassoon…
	Defilippi et al., 2019		pro	-	yes	piano, guitar, bass, accordeon, violin, sax..
	Shahin et al., 2008		pro/ amateur	-	yes	violin, piano
Group 3 - Test	WING	Delogu et al., 2006.	-	-	-	-
		Slevc & Miyake, 2006	-	-	-	-
	Seashore	Milovanov et al., 2010	-	-	yes	voice
	Golds MS	Han et al., 2019	8	0	(2 yrs ago)	-
	AMMA	Coumel et al., 2019	amateur	0	-	voice
		Elmer et al, 2012	pro	7	-	flute, piano, violin, cello

A.2 – Chapter 2

Table 2

Consonant productions

	Production 1	Production 2	Production 3	Production 4
/aka/				
duration	318 ms	303 ms	382 ms	328 ms
/apa/				
duration	321 ms	310 ms	283 ms	269 ms
/asa/				
duration	410 ms	343 ms	408 ms	325 ms
/ata/				
duration	277 ms	272 ms	261 ms	375 ms

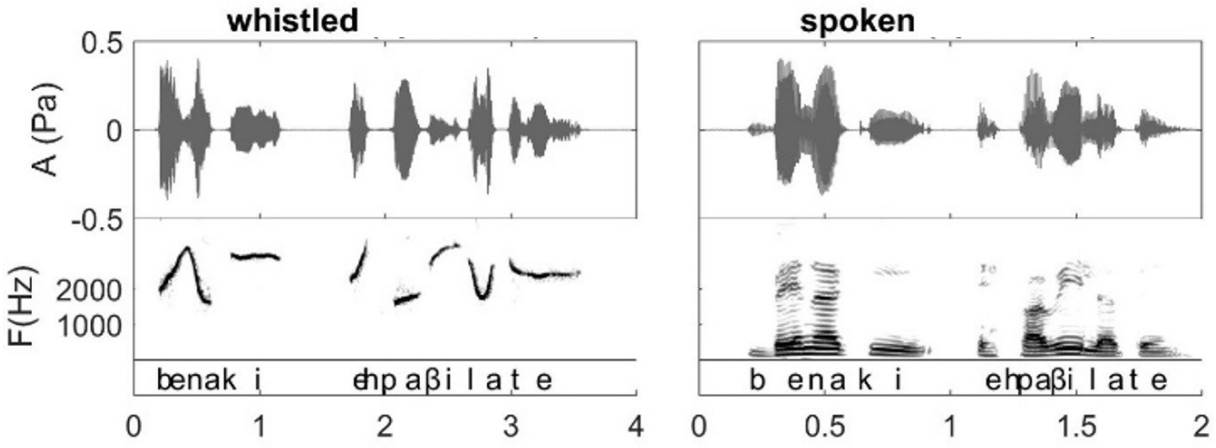


Figure 1 of SuppPub1: Waveforms and Spectrograms of the Spanish sentence "ven aqui espavilate" expressed both in spoken and whistled forms by the same whistler (transcribed in phonetic transcription and meaning "come here, hurry up"). In this figure, we can follow the dynamics of the sentence in both modalities. It illustrates the spectrographic differences/similarities between spoken and whistled modalities that are explained in the introduction of the paper. Vowels are distributed at different frequency levels and consonants represent modulations in frequency and amplitude of the more steady frequencies of the vowels. (Recording and editing by Julien Meyer, listen to sound extracts in https://soundcloud.com/user-28976943/figure2-sentence-speaker1-1?in=user-28976943/sets/meyer-and-diaz-2021-sounds-of-whistled-speech).

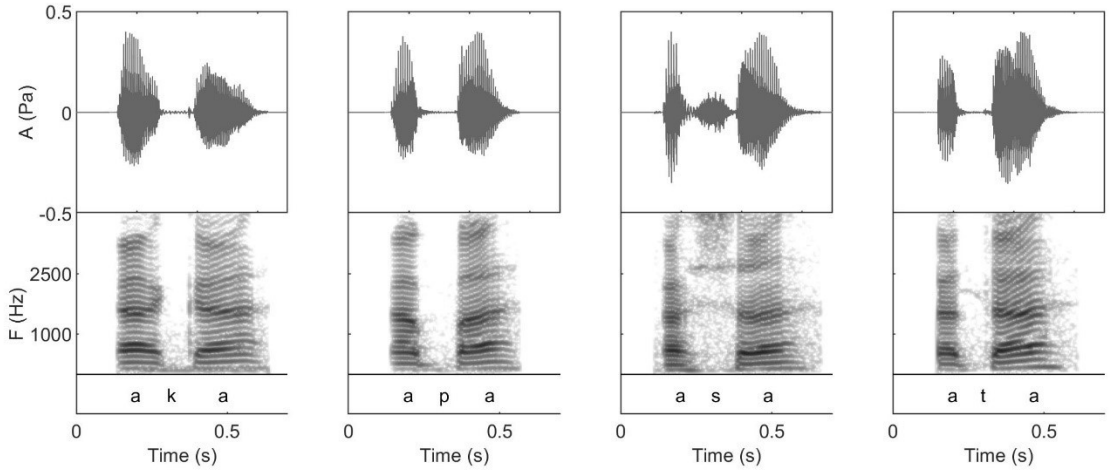
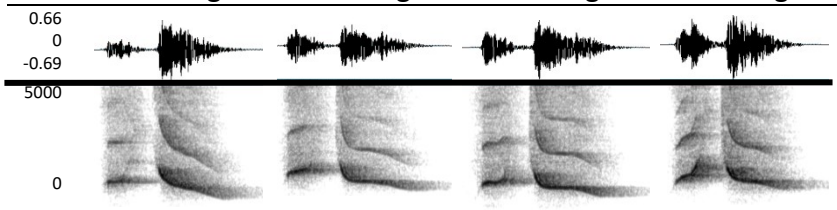
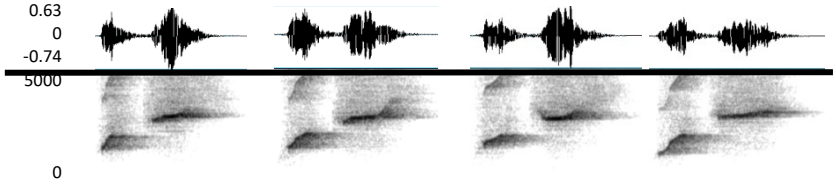
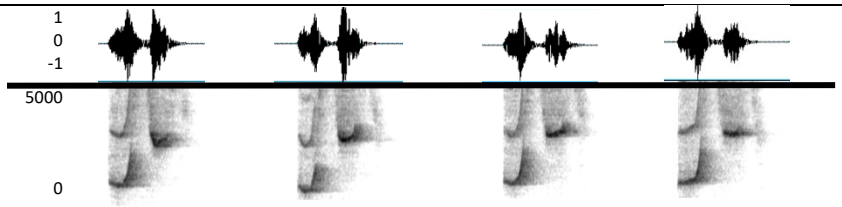
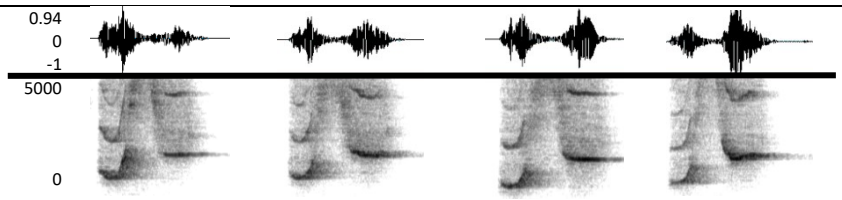
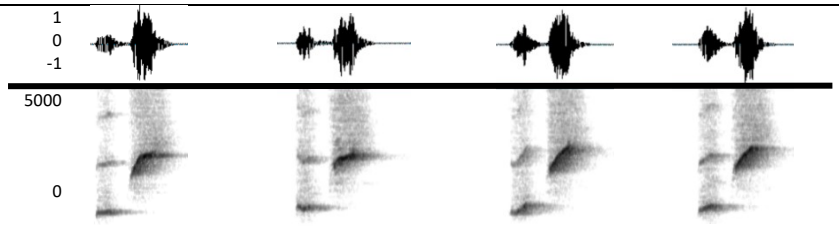
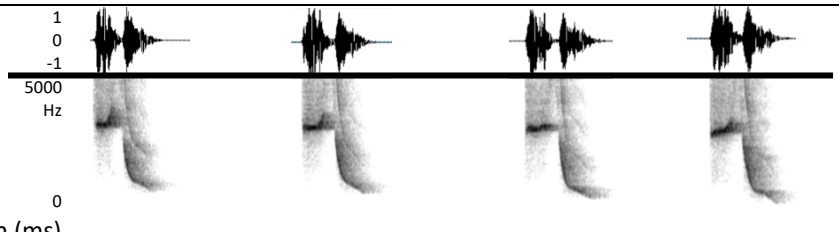


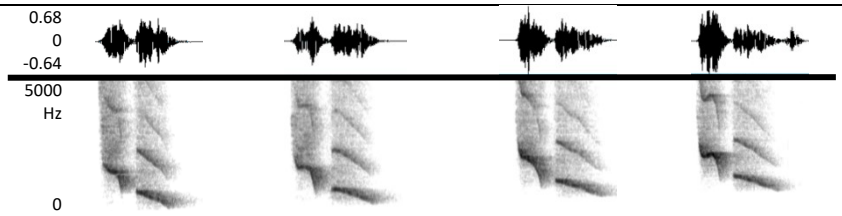
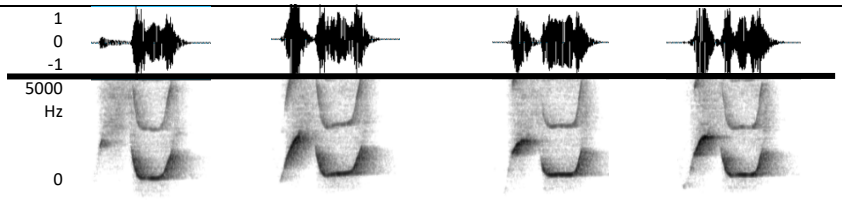
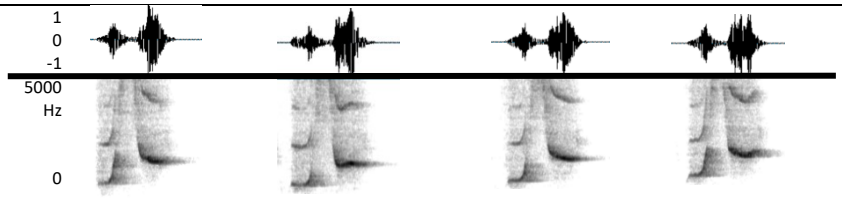
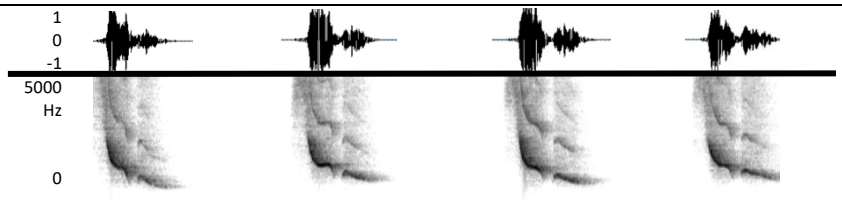
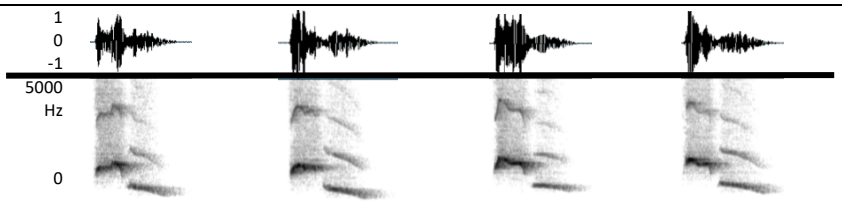
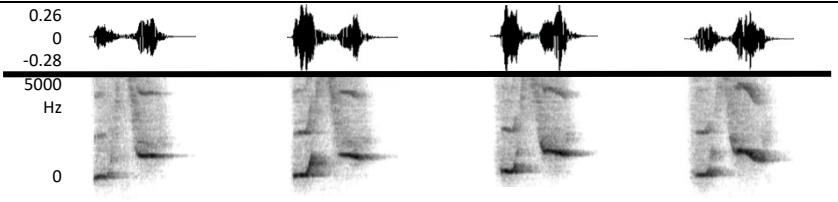
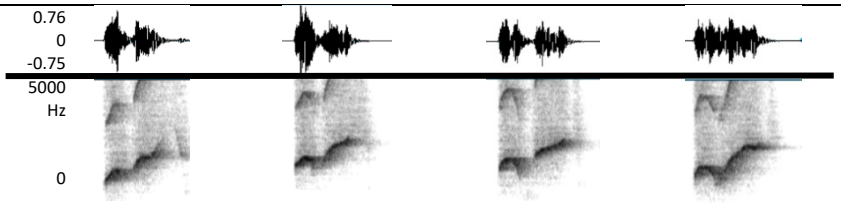
Figure 2 of SuppPub1: Waveform and Spectrogram of the **spoken Spanish VCV** forms corresponding to the whistled VCV forms of the experiment (Recording by Julien Meyer and we thank Laure Dentel for help in editing)

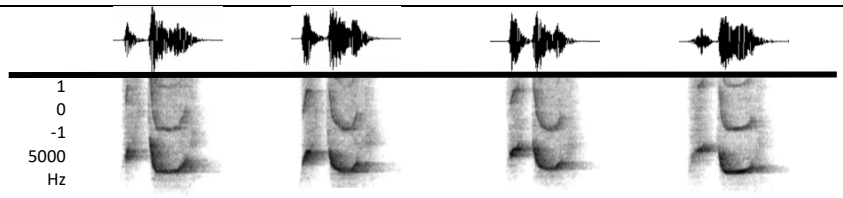
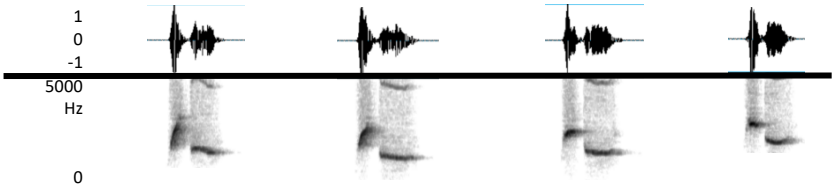
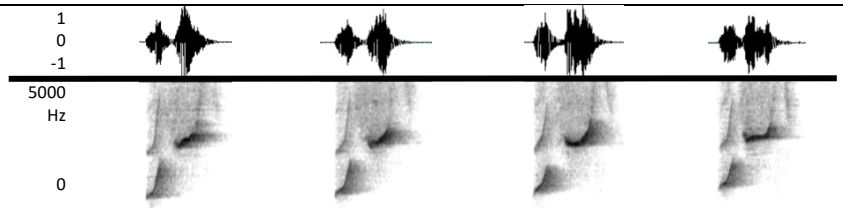
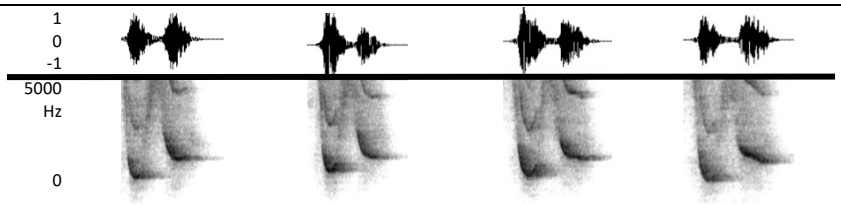
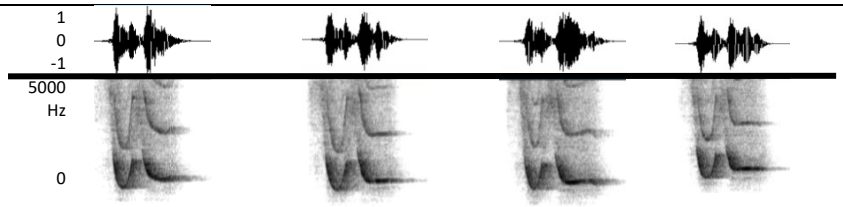
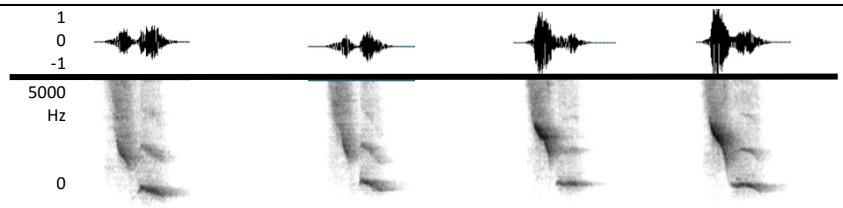
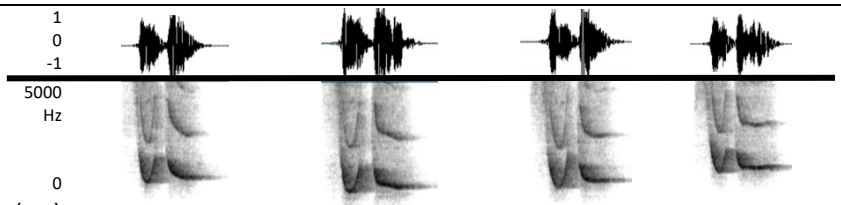
A.3 – Chapter 3

Table 3

Whistled word stimuli descriptions

Word	Spectrogram				Comments
	Recording 1	Recording 2	Recording 3	Recording 4	
/bato/					Slo amplitude rise, frequency rise and interruption due to the consonant /t/, /o/ lowers (Recording 1)
Duration (ms)	860	960	990	832	
/bekij/					Slow amplitude rise before /e/ and interruption jump to /i/, effect of /j/ most noticable on Recordings 2 and 3
Duration (ms)	600	690	650	750	
/kasis/					Small amplitude rise for /k/.Slight descent in pitch from /k/ to /a/ and fast rise from /a/ to /s/. Pitch lowers very slightly for /i/, final /s/, slight fq rise for Recs 1, 2 and 3
Duration (ms)	800	770	850	820	
/koʃe/					Small amplitude rise /k/, slight descent to /o/, fast rise to /ʃ/, attaining a higher pitch and landing on stable /e/ (Rec. 4 descent from /ʃ/).
Duration (ms)	837	933	933	869	
/kopi/					Small amplitude rise for /k/ on low and stable /o/. Cut to the second syllable, pitch rise towards /i/ begins with the /p/ (see rec 1, 3 & 4), unstable for rec 3 4.
Duration (ms)	680	758	727	696	
/kilo/					Large amplitude rise for /k/, Stable /i/ with a small rise to /l/ (recordings 1, 2 and 3), before descending to /o/ with no interruption
Duration (ms)	799	845	876	768	

/depo/						Descent from a higher point towards short and unstable /e/. Short but clear interruption before /po/. Final /o/ descends slightly
Duration (ms)	925	1035	1066	1035		
/final/						Disyllabic but almost 3 syllables as the final rise towards /l/ creates a change in frequency
Duration (ms)	1039	1100	1070	979		
/fose/						Short and stable /o/ pitch before rise to /s/. Sharp descent into /e/ (recordings 1 and 2), or no descent (rec 4).
Duration (ms)	929	907	923	989		
/ʃamo/						Continuous frequency descent, starting with an almost spoken /ʃ/ short /a/, very slight interruption, where /o/ descends slightly afterwards
Duration (ms)	896	1010	1020	1060		
/mego/						Large amplitude rise for /m/, slight pitch augmentation with /m/ into /e/ and then a cut and to stable /o/.
Duration (ms)	910	940	950	940		
/pase/						Slow amplitude rise /p/ (Rec 1 and 4), low stable frequency for /a/, rise towards /s/ and then slide down to/e/.
Duration (ms)	816	877	920	859		
/peʁil/						Airy /p/ with a pitch rise towards /e/ followed by a slight dip in frequency. Then pitch rise towards /il/
Duration (ms)	750	726	812	757		

/petaʁ/					/p/ articulation (slow amp rise) rises in pitch with /e/ towards /t/. Short interruption before descent to /a/. Final rise towards /ʁ/, can create the illusion of 3 syllables.
Duration (ms)	816	877	920	859	
/pike/					Sharp rise up to/i/ (particularly for recordings 1 and 2) and cut to stable /e/
Duration (ms)	728	800	823	744	
/polis/					Slow amp rise for /p/, continuous pitch change, passing through /o/, rising to /l/, to a stable /i/ (Rec 4), before ascending to the /s/.
Duration (ms)	751	794	786	794	
/saʃe/					First pitch descent to /a/, then rising higher to /ʃ/ and finally descending to stable /e/ (Rec 1, 2 & 3)
Duration (ms)	859	843	987	885	
/sona/					Fast pitch descent to short /o/, before a rise in frequency towards /n/, with a final descent to stable /a/
Duration (ms)	919	892	963	901	
/siʁo/					Descending fricative, barely pausing on /i/, until reaching stable /o/. We hear a cut for Recordings 1 and 2, but not for 3 and 4
Duration (ms)	772	788	683	756	
/soda/					Pitch starts high, descends to very short /o/, rise to /d/, w/interruption, and descent to stable /a/
Duration (ms)	822	910	848	980	

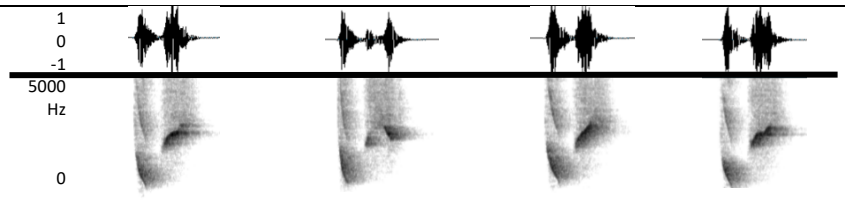
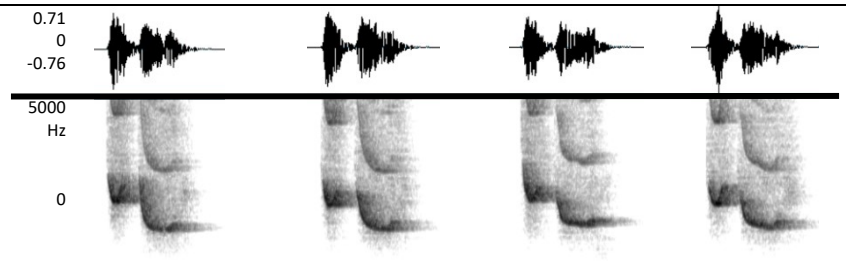
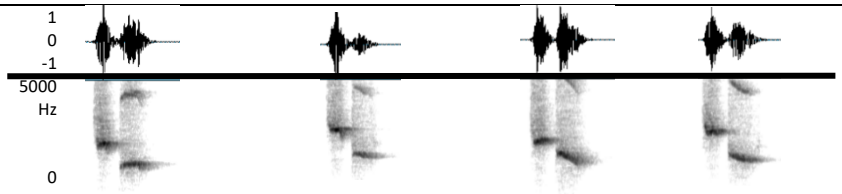
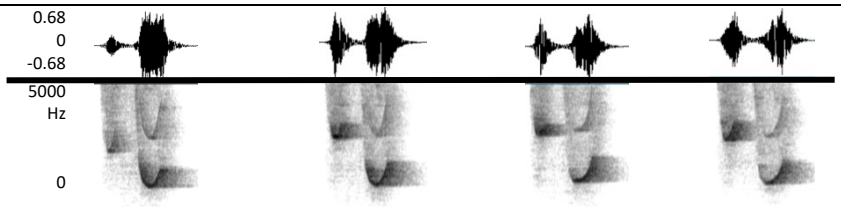
/tapi/						Initial /t/ articulation starts high pitch, descends to /a/ without stopping. Slight interruption, and rise continuously to /i/
Duration (ms)		646	709	656	700	
/tetaʁ/						High /t/ pitch, descends towards very short /e/ which rises again to /t/ before descending to a stable /a/. Slight rise towards /ʁ/ (especially for rec 1, 2 and 3).
Duration (ms)		730	713	705	738	
/tike/						Amplitude rise for /t/ and /k/, clearly articulated. Stable vowel pitches for /i/ and /e/
Duration (ms)		653	636	670	645	
/tizan/						High frequency for /t/ descends to /i/, rises towards /z/ and the descends to a very short /a/, which rises again to /n/, almost producing three syllables
Duration (ms)		703	732	717	761	

Table 4:

Spoken Word Descriptions

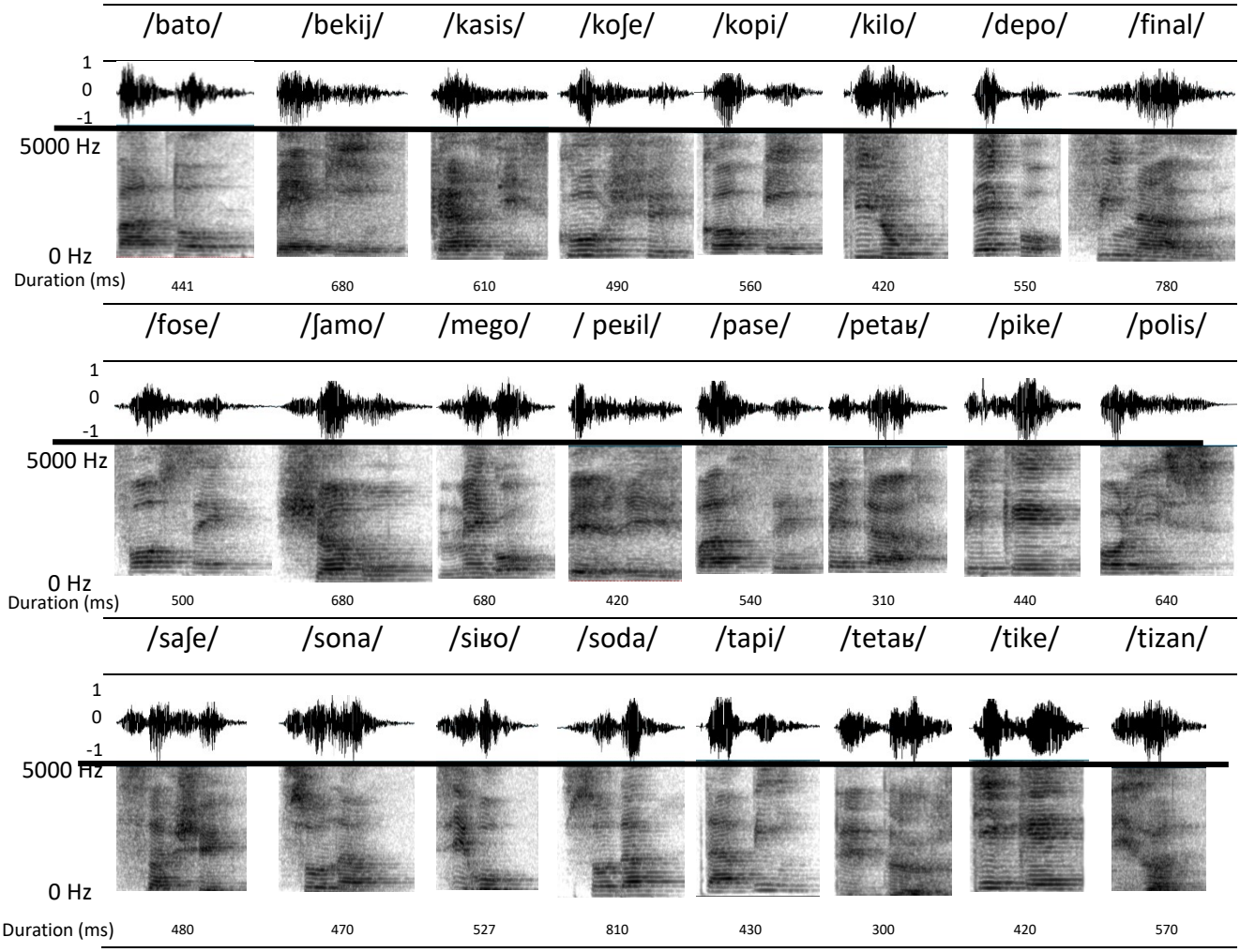


Figure 2:

Duration of whistled and spoken words

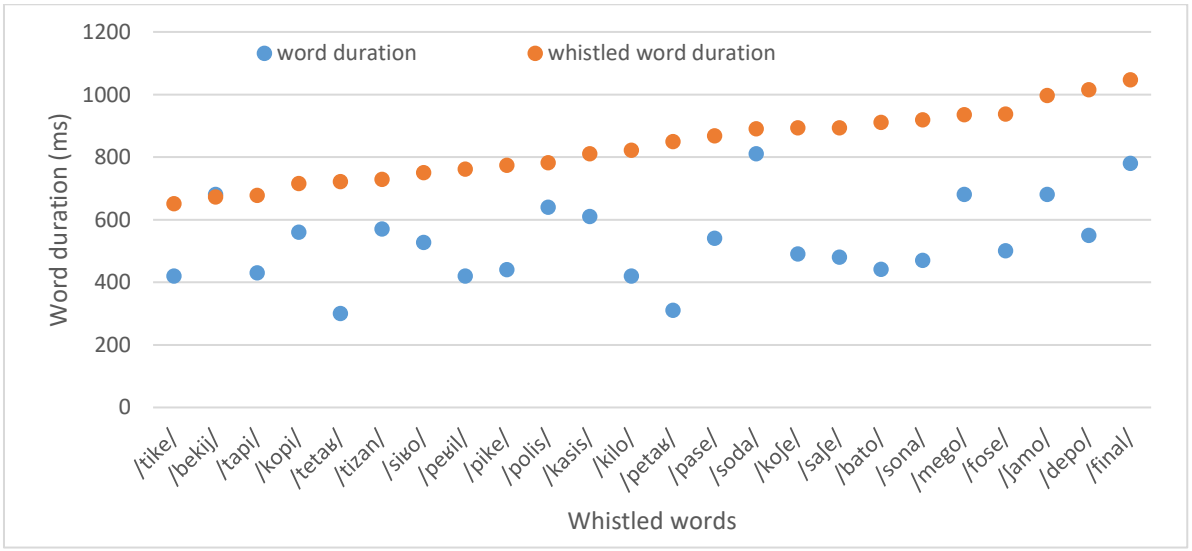


Figure 3:

Variability in duration for whistled word recordings

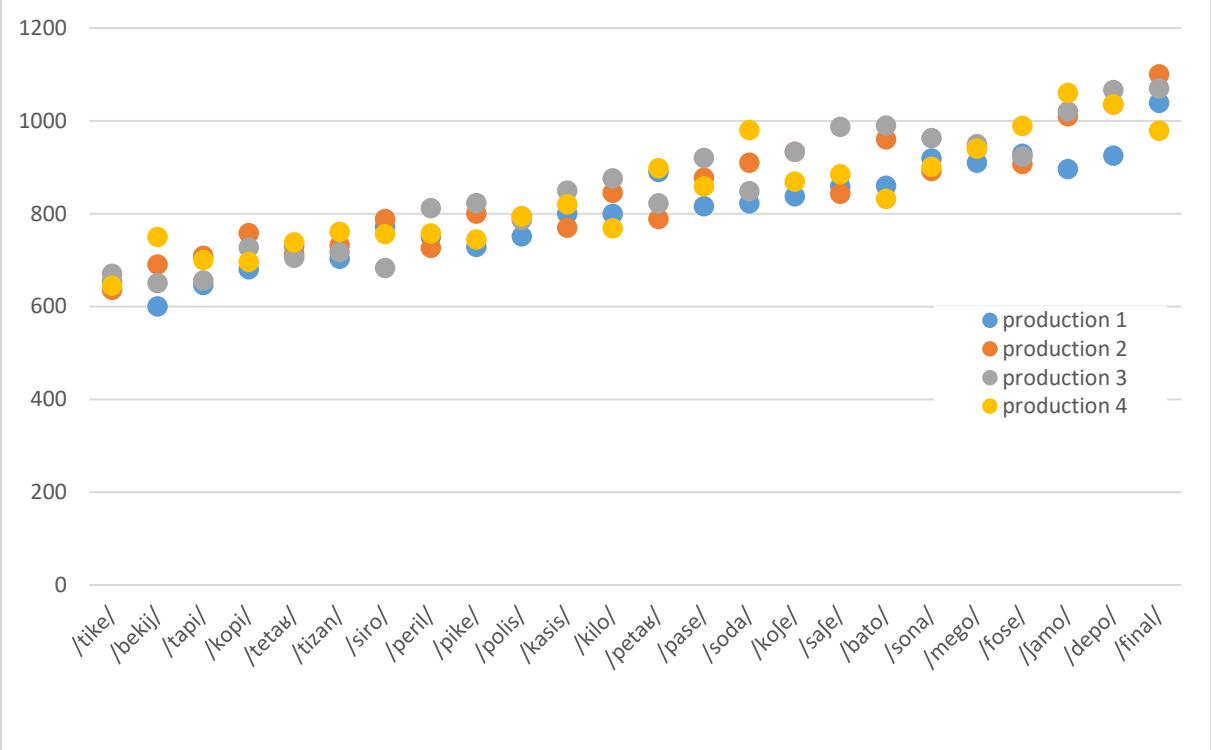
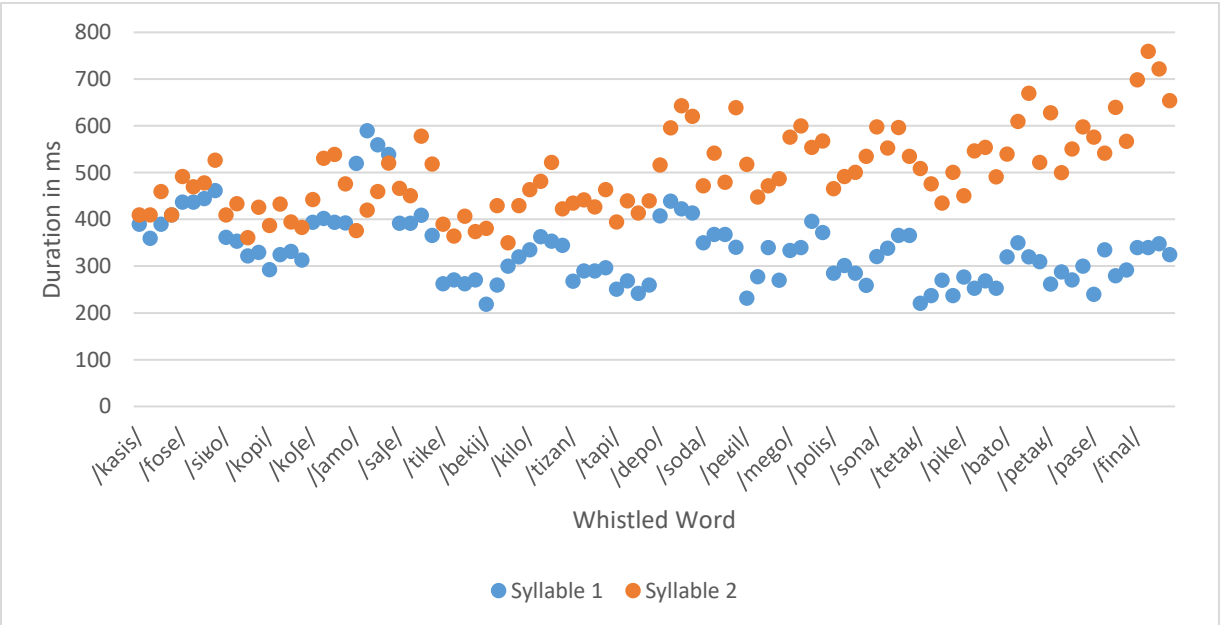


Figure 4:

Comparison between S1 and S2 syllable duration for whistled words according to difference in ms



A.4 – Chapter 4

A.4.1 Reflections on participant comments (phonemes)

To further understand participants' categorization methods as well as their own awareness of the cues present in the whistled phonemes, we analyzed the questionnaires proposed at the end of the vowel and consonant experiments (Chpt. 4). The questionnaires included the following questions: "Do you feel that the training portion (where we indicated whether you responded with the correct or incorrect answer) helped you to recognize and categorize these vowels (consonants)? Why or why not?[24]", "Describe the method used to identify the vowels (consonants) before and after the training portion. Were certain vowels (consonants) easier to recognize?[25]"

A.4.1.1 Vowels

Training effect

There were 64 questionnaire responses obtained for the vowel portion of the experiment. This included 30 participants among the "non-musicians" (levels 1,2, and 3) and 34 among the "musician participants" (levels 4, 5, and 6). Among these responses, participants generally found the training portion helpful (66.6% of non-musicians compared to 63.3% of musicians), with only a 3.3% difference between the two groups. Some of the critiques concerning the vowel experiment indicated that it was "too short", or "unhelpful", and several

[24] *« Est-ce que l'entraînement (où nous avons indiqué si vous aviez répondu avec la bonne ou mauvaise réponse) vous a aidé à reconnaître et catégoriser les voyelles (les consonnes)? Pourquoi, ou pourquoi pas? »*

[25] *« Décrivez la méthode que vous avez utilisée pour identifier les voyelles (les consonnes) avant et après l'entraînement. Est-ce que certaines voyelles (consonnes) étaient plus faciles à reconnaître? Pourquoi? »*

participants indicated that this was because they felt like they intuitively knew the vowels. Among the musician participants, 4 of the comments mentioned the interference of absolute pitch, where these participants heard note names rather than a vowel. Some participants also mentioned having difficulty in part 3, because of the increased amount of variation.

Categorization strategy and vowel preferences

The following categorization strategies were mentioned in the vowel responses overall: pitch, timbre, thinking of the vowels (or intuitively knowing which vowel corresponds to which sound), duration, and evolution in the sound. Participants often described using multiple strategies for categorizing these vowels. Descriptions of timbre included “rounded”, “strident”, “dry”, and “round”. Interestingly, we could also contrast participants who mentioned intuitively knowing the vowel categories with other participants who indicated that they had a lot of difficulty with this part of the phoneme experiment.

When asked to indicate vowel preference (or easiest to categorize), participants systematically indicated /i, o/ and /i/ and /o/. One musician participant indicated finding /a/ easiest to categorize and was able to use /a/ to place the other vowels. However, this seemed to be an exception, as none of the other participants mentioned finding /a/ or /e/ easy to categorize. Some participants (both musicians and non-musicians) also indicated the vowel order (i, e, a, o) from highest to lowest. Generally, the responses given by the participants were very similar, regardless of their musical level (see Table 1), though non-musician participants mentioned using pitch more often than musician participants (even if one musician participant described the relationship between the vowels as the second inversion of a major chord). Interestingly, 6 musician participants indicated that they did not succeed at

the task (17.6%), however, as several participants mentioned confusion due to absolute pitch in the comments, it may be important to consider this factor in further whistled speech tasks.

Table 1*:*

Vowel categorization method and preferred vowels for non-musicians and musician

	Categorization method				
	Pitch	**Thinking of vowels**	**Timbre**	**Duration/evolution**	
Non-musician	63.3%	20%	10%	10%	
Musician	47.1%	20.6%	14.7%	11.8%	
		Preferred vowels		**Full order**	**Just i/o**
	/i,o/	**/i/**	**/o/**		
Non-musician	33.3%	13.3%	10%	23.33%	16.7%
Musician	26.5%	2.9%	5.8%	17.6%	5.9%

A.4.1.2 Consonants

Training effect

In quantifying the results obtained from the questionnaires for the consonant experiment, we can consider the responses of high-level musician participants separately from participants with no musical experience or low levels of musical experience. Only 27 participants in the non-musician and low-level musician group, responded to the first question, whereas 36 participants responded within the high-level musicians.

The majority of both of these participant groups indicated that the training was helpful: 62.1% of none and low-level participants indicated that the training was helpful, as did 82.4% of high-level participants. According to participants' comments, this was often because it allowed participants to "confirm", "differentiate" and "memorize" the whistled consonants. However, we note that 27.6% of the none and low-level participants commented negatively

on the training portion. This was more often than the high-level musician participants, as only 17.6% did so, either indicating that it was unhelpful, or that it was helpful despite being "too short", or "too fast". Often, participants mentioned that the training portion should have included the correct responses. Interestingly, some participants specifically cited the consonants that they were able to differentiate due to the training portion, this included /s/ and /p/ (for 4 participants), /k, p/ for 2 participants, and /k, t/ for 2 participants.

Consonant preferences and categorization methods

The second question asked participants to specify which method they used to categorize consonants, and if some were easier to recognize. Participants mentioned several different strategies. These include the use of pitch to distinguish certain consonants or associating the whistled sound heard with the sound of the consonant ("hearing the consonant", which we will refer to as "thinking of consonants"). Participants also mentioned using the duration (or the comparison between segment duration), the articulation (either using the term "articulation" in their descriptions, or by associating attacks to consonants and describing the amount of air used), and timbre qualities which were described with adjectives such as "soft", "dry", or "strident" (with slightly more descriptive terms used by the musician participants). Participants often described consonants using a combination of these different strategies.

Participants also indicated their preferred consonant, highlighting differences between the non-musicians and low-level musicians and the high-level musician responses. Indeed, the first group of participants mentioned a preference for /s/, as well as for the pair /s, t/, see Table 2. The high-level musician participants declared preferring pairs more often than individual consonants, with a preference for /s, k/, followed closely by /s,t/ and /k,t/ (see Table

3). These preferences, as well as further comments, underlined an awareness of the opposition between the different groups of consonants, grouping either the articulatory cues (/k,t/ and /s,p/) or the frequency change (/s,t/ and /k,p/). In these responses, participants mentioned pitch cues most often, followed by articulation. In addition, we observe an awareness of the difficulty presented for /p/, which was most difficult to categorize.

Interestingly, some comments described the whistled consonants very differently from the acoustic cues in the signal. This is especially the case for participants who mentioned "thinking of the consonant", sometimes describing the consonant as a "*'doubled' whistled sound", a "t-sound" or a "k-sound".* This underlines difficulties in describing the cues despite using them in the categorization task. This also differentiates the group of non-musician and low-level musician participants from the high-level musicians. Indeed, through their descriptions, the high-level musician participants indicate an awareness of pitch, interruption, and sometimes articulation.

Table 2:

None and Low-level participants categorization methods and consonant preferences

Categorization method					
Pitch	**Thinking of consonants**	**Duration**	**Articulation**	**Timbre**	
43.3%	13.3%	20%	20.69%	20%	
Preferred consonants				**Groupings**	
/s/	**/k/**	**/p/**	**/t/**	**/k,t/&/s,p/**	**/s,t/&/k,p/**
36.7%	3.3%	0%	6.7%	3.3%	16.7%
/s,t/	**/s,k/**	**/s,p/**	**/k,t/**	**/s,k/&/t,p/**	
16.7%	3.3%	0%	3.3%	0%	

Table 3:

High-level musician participants' categorization methods and consonant preferences

Categorization method					
Pitch	**Thinking of consonants**	**Duration**	**Articulation**	**Timbre**	
44.4%	22.2%	5.6%	38.9%	22.2%	
Preferred consonants				**Groupings**	
/s/	**/k/**	**/p/**	**/t/**	**/k,t/&/s,p/**	**/s,t/&/k,p/**
5.6%	2.7%	0%	2.7%	13.9%	16.7%
/s,t/	**/s,k/**	**/s,p/**	**/k,t/**	**/s,k/&/t,p/**	
16.7%	22.2%	0%	16.7%	5.6%	

A.4.1.3 Discussion

Through these comments following the two phonological categorization tasks, we consider that participants show an awareness of phonological cue categorization which reflects their behavior in these tests.

Indeed, the vowel and consonant preferences described correspond to the hierarchies found in the statistical analysis. In the vowels, this includes the opposition between /i/ and /o/ (and sometimes the vowel order). We also observe how musicians and non-musicians describe using similar tools for categorization, though the musicians sometimes mention interference from musical experience.

In consonants, we also observe a parallel between the descriptions of performance and the statistical results. This applies to participants' preference for /s/ and /t/, and the distinction of consonant pairs which regroup consonants according to defining cues. In

addition, participants mention these cues specifically, especially pitch, which generally suggests an awareness of such cues. We wonder how important the ability to describe acoustic cues is for categorization, as pitch (or frequency) is one of the only cues to be regularly described. The ability to describe the cues present is one notable and major difference between musicians and low-level/non-musician participants. The ability to describe and therefore categorize sounds may help with the categorization task at hand. We also observe a shift in consonant preferences, where /k/ is mentioned more often for high-level musician participants, which also holds true with statistical performances.

However, in both the whistled vowel and the whistled consonant task, we also notice the strong subjectivity of perception. This is the case concerning the effect of the training in both the whistled vowel and the whistled consonant experiment. In terms of the whistled vowels, we notice that 63-66% of participants indicated that the training was helpful, even though statistically, only some musician participants show a learning effect. In the consonants, this rate is at 62% for none and low-level participants (who showed no learning effect) and 82% for high-level musicians, who did show a learning effect. Thus, though some participants correctly describe their behavior, participants are not always aware of their behavior. Some examples include participants declaring the training portion shorter in either the vowel or consonant experiment (when comparing the two). Others indicated that the number of times each phoneme appeared seemed unbalanced (which was not the case, due to the fixed design of the experiment). Thus, although these comments provide insight into participants' behavior, adding to the statistical results by showing how participants think they perceive these whistled speech sounds, they are not fully accurate. This suggests that participants are not entirely conscious of their behavior in these tasks.

A.4.2 Isolated whistled vowels and the effect of musical instrument

A.4.2.1 Introduction

In Chapter 4.1, we find an effect of musical expertise on the vowel, notably for vowels /e/, /a/, and /o/, more specifically present for whistler B (notably for /e/ and /o/). We also concluded that musician participants were more strongly affected by the whistler heard and the larger range. In Chapter 4.2, we specified that musical advantage depended on the level of musical expertise and more specifically, the effect of musical instrument. Thus we wonder if elements such as whistler range would be a cue used by all musicians, regardless of musical experience, or specific to instrument specialization.

A.4.2.2 Method

We used an identical method to the experiment presented in Chapter 4.1.

Participants

As we included only musicians with a high level of instrumental expertise, we reduced the 36 musician participants included in Chapter 4.1 to 32. This only includes participants who specialize in the piano, the violin, in voice or in the flute. As we also included the 30 non-musician participants, thus there are 62 total participants in this part of the analysis.

A.4.2.3 Results

We applied a GLMM on Correct Answers, with Part (P1, P3), Vowel (/a/, /e/, /i/ and /o/), Whistler (A, B) and Instrument (Violin, Piano, Voice and Flute) as fixed factors, and Participant as a random factor. We find a significant effect of Whistler (X^2 (1, *N*=62) = 46.263,

p <.001), Vowel Played (X^2 (3, *N*=62) = 247.866, *p* <.001), Part*Vowel (X^2 (3, *N*=62) = 16.755, *p* <.001), Whistler*Part (X^2 (1, *N*=62) = 5.442, *p* =.020) and Whistler*Vowel (X^2 (3, *N*=62) = 17.30, *p* <.001).

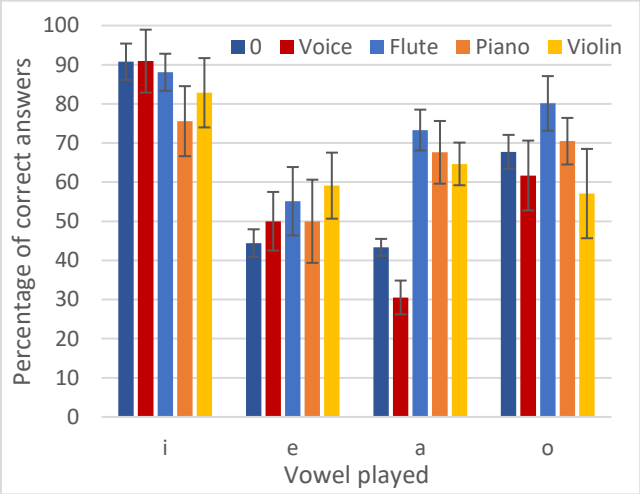
We also find several interactions including the factor instrument. This was present between Instrument*Vowel, as well as a number of significant double interactions: Whistler*Vowel*Instrument, Part*Vowel*Instrument, and a triple interaction Whistler*Instrument*Vowel*Part. As we previously analyzed these effects according to the factors Musician/Non-musician (see Chpt.4.1), and because this group includes largely the same participants, we will focus on the interactions found which include the factor Instrument. We analyze the significant interactions with post-hoc tests with the Bonferroni correction. When considering the double interactions, we once again decomposed the results by applying a GLMM to each vowel played.

The significant interaction Instrument*Vowel (X^2 (12, *N*=62) = 111.747, *p* <.001), was analyzed with a post-hoc test. This shows significant differences between certain instruments and non-musicians, as well as within the Instruments for the vowel /a/. Indeed, we find a significant difference where Flute > Non-musicians for the vowel /a/ *(p* =.009), and a tendency between Non-musicians and Piano (*p* =.07) where we find that Piano > Non-musicians.

Between the different instruments, we also find significant differences for the vowel /a/ where Flute < Voice (p <.001), Piano > Voice (p =.005) and Violin > Voice (p =.012), see Figure 1.

Figure 1:

Percentage of correct responses obtained per vowel



We also obtain different hierarchies for each of the instruments which differ from that of non-musicians where /i > o > a = e/, see Table 1.

Table 1*:*

Vowel hierarchies according to Instrument

	Significant differences		**Vowel hierarchies**
Voice	/i > o/, /i > a/, /i > e/, /o > a/	(*p*s < .001)	/i > o (=e) > a (=e)/
Flute	/i > e/, /o > e/	(*p*s < .001)	/i = o (= a) > e (=a)/
Piano	/i > e/	(*p* < .001)	/i = o (= a) > e (=a)/
	/o > e/	(*p* = .02)	
Violin	/i > a/	(*p* = .02)	/i > o = a = e/
	/i > e/, /i > o/	(*p*s < .001)	

To analyze the double interactions, Whistler*Vowel*Instrument (X^2 (12, *N*=62) = 65.220, *p* <.001), Part*Vowel*Instrument (X^2 (12, *N*=62) = 40.176, *p* < .001) and the triple interaction Whistler*Instrument*Vowel*Part (X^2 (12, *N*=62) = 42.154, *p* < .001), we separated the analysis according to Vowel (/i/, /e/, /a/ and /o/). We then applied a GLMM on Correct Answers, with Part (P1, P3), Whistler (A, B) and Instrument (Violin, Piano, Flute, Voice) as fixed factors.

We find several significant interactions within the vowel /i/ which include Instrument: Whistler*Instrument (X^2 (8, *N*=62) = 25.342, *p* =.001) and Instrument*Part (X^2 (5, *N*=62) = 16.351, *p* =.006). When applying a post-hoc test to the interaction Whistler*Instrument, we observe significant differences between non-musicians and piano for Whistler B (*p* <.001), where non-musicians > pianists. In our analysis of Instrument*Part, we find no significant effects within the post-hoc test.

We find several significant interactions for /e/ which include the factor Instrument: Whistler*Instrument (X^2 (8, *N*=62) = 39.2, *p* <.001), Part*Instrument (X^2 (5, *N*=62) = 16.6, *p* =.005), and a double interaction Whistler*Part*Instrument. (X^2 (5, *N*=62) = 11.8, *p* =.026). We applied a post-hoc test to Whistler*Instrument, where we find a significant difference for pianists for whistlers, where A < B (*p* <.001). In the application of a post-hoc to Part*Instrument, we find a significant difference between parts 1 and 3 for violinists (*p* =.011), where 3 > 1.We applied a post-hoc test to the double interaction Whistler*Part*Instrument which shows a significant difference between whistlers A and B, for pianists in part 1 (*p* =.011), where B > A.

For the vowel /a/, we find only one significant effect which includes Instrument, and that is the Main effect of Instrument (X^2 (8, *N*=62) = 44.08, *p* <.001). Because of this, the results are identical to those of the significant interaction Instrument*Vowel.

For the vowel /o/, we find a significant interaction between Whistler*Instrument, X^2 (8, *N*=62) = 23.30, *p* =.001). We applied a post hoc to this interaction, where we find a significant difference between whistlers A and B for voice (*p* =.012) and for piano (*p* =.002), where B > A.

A.4.2.4 Discussion

In our analysis of instrument specificity, we were able to further detail the musician advantages presented in Chpt.4.1, where we confirm that the advantages for /a/, /e/, and /o/ can be specified according to instrumental expertise. We observe various profiles according to the instrument specialization in interaction with the vowel, which first suggests that instrumentalists do not behave in the same manner according to expertise.

This is especially the case for /a/, where flutists are the only instrument to show significant differences with non-musicians. However, we also observe significant differences among instrumentalists, as flutists, violinists, and pianists show significant differences with singers for /a/. Such differences in profiles are also reflected in the vowel hierarchies shown for each instrument group. We further explored these differences when analyzing the vowels individually. This shows that the the advantage of non-musicians over musicians for the vowel /i/, can be attributed to pianists' performances more specifically. In addition, we observe a learning effect for /e/ which is specific to violinists. In Chapter 4.1, we also observed an advantage for whistler B over whistler A for vowels /e/ and /o/. Here, we specify these

differences according to instrumental expertise, notably pianists for /e/, and both pianists and singers for /o/.

These differences suggest that some of the musician effects are common to all instrumentalists (the advantage for /e/ and for /o/), however others are specific to a single instrument group (notably the advantage of the wider vowel range, i.e. whistler B). We therefore confirm the findings shown in Chpt.4.2, showing the specificity of instrument specialization for whistled vowels. We also underline an advantage for flutists, which, as suggested in Chpt.4.2, could be attributed to similarities in timbre between the flute and whistled speech. The clear advantage for /a/ (for flutists), and disadvantage for /a/ (for singers compared to other musicians) is particularly interesting. We wonder if singers are affected by the treatment of music-like sounds as speech. Indeed, as singers adapt musical notes to various vowel sounds the association between whistled speech and a single vowel may be difficult. We could also consider the possible impact of absolute pitch, as mentioned by participants in A.4.1.

A.5 – Chapter 5

A.5.1 Reflections on participant comments (words)

A.5.1.1 Introduction

Following their participation is the whistled word experiment (Expt 8), participants responded to a short questionnaire, asking them to indicate if they had used a certain method to identify the whistled words and if certain words were easier to identify than others. To quantify the responses given by individual participants, the most common categorization methods or remarks mentioned will be described here, as well as the number of participants who wrote about these themes. Out of the 93 participants, only 89 participants responded to the questionnaire, and with varying amounts of detail. This includes 17 participants in level 0, 10 participants in level 2, 23 participants in Level 3, 15 participants in level 4, 6 participants in level 5, and 18 participants in level 6. We considered participants' responses according to these different levels, allowing us to differentiate these self-declared musical skill sets and how such differences might affect whistled speech perception.

A.5.1.2 Quantifying questionnaire responses

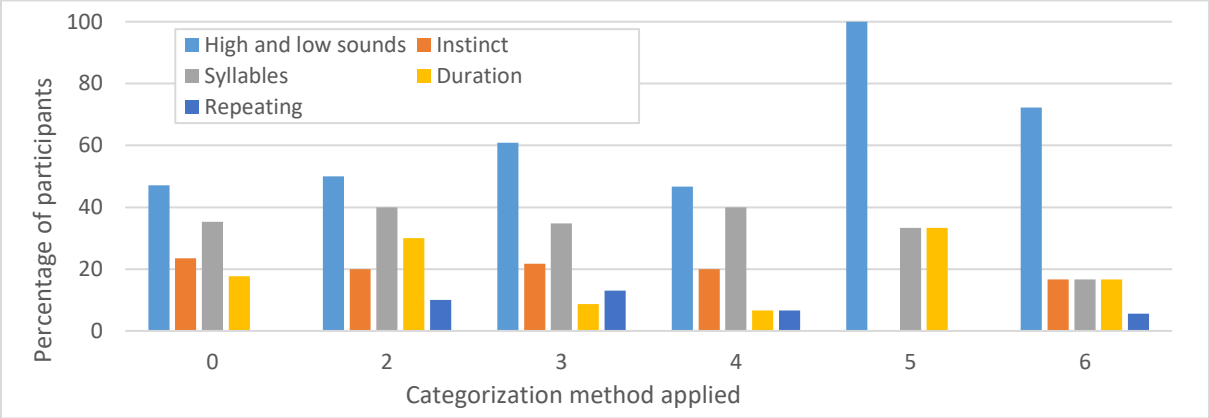
Identification method

In each of the musical levels, participants mentioned that this was a difficult task. An average of 20% of participants in each level of musical experience made this remark: 0 – 23.53%, 2 – 20%, 3 – 13%, 4 – 26.67%, 5 – 16.66%, and 6 – 22.22%. Five participants even indicated that the experiment was simply too difficult to understand or apply any method whatsoever. The most commonly mentioned methods used to categorize whistled words

include the mention of “high” and “low” sounds (i.e. pitch contrasts) within the whistled words (see Figure 1). Other methods included using the syllables present, the duration of the word, and one’s “instinct”.

Figure 1:

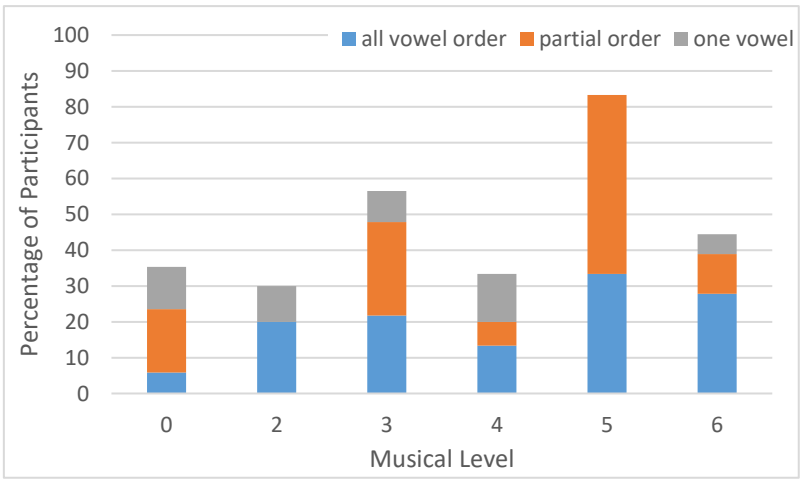
Percentage of participants using cited identification methods according to Musical Level



Among these responses, 34.7% of participants gave a partial or complete order of the whistled vowel pitches. This was especially the case for participants in levels 3 and 5 (see Figure 2). In cases where participants only highlight the recognition of a single vowel, this often corresponded to /i/. The word preferences mentioned included /kilo/, /kopi/, /polis/, or even /tapi/, further reiterating the opposition between “high” and “low” vowels.

Figure 2:

Percentage of participants indicating vowel order



The second most frequently mentioned method is using the syllabic structure, which participants often linked to duration. Participants sometimes described trying to match the syllables with the pitch changes present (which they described as being rather difficult or confusing, especially for words such as /tizan/ or /final/) or using the number of syllables and the duration to identify words. Interestingly, participants with less musical experience indicate using whistled syllables more than participants with higher musical levels (for example 5 or 6). Some participants described having whistled the words themselves, or trying to imitate the whistle and attempting to match these frequencies to the "melody" of the spoken word.

Although these responses mainly focused on the vowels of the whistled words, a few participants also described the consonants. This was especially the case for participants with musical experience above Level 3 (only one participant in Level 2 described consonant movements). These descriptions vary in consonant preferences: some participants underline that the consonants /k/, /p/, and /t/ as being easier to categorize, whereas others indicate that fricatives, or /ʃ/ and /s/ are simpler (specifically mentioning the "/s/ in /polis/"). They associate these fricatives with "glissandos", "turns" and "slides", whereas the plosives are described as "dry", "percussive", "short" or "clear". Two participants also described linking the whistled words to their semantic value. Finally, several participants indicated not using a method and simply relying on their intuition.

Describing whistled speech

The responses proposed by participants with varying levels of musical experience provided another level of insight into how whistled speech was perceived by listeners. Interestingly, participants with a higher level of musical experience often described whistled

speech in a systemized and almost pedagogical way (despite not having encountered this form of modified speech before). We can cite the following examples[26]:

> *"After a few words (maybe around 10), I started to associate pitches with vowels (high for "i" and low for "o"), as "i", "é" and "è" bring out the higher harmonics more than the vowels "a", "e" "ou" and "o". After another 10 words, I started hearing what the consonants could represent (with the same logic as the vowels) ….I think I ended up understanding certain consonants ("ch" and "s" for example) because they had brief pitch changes between two vowels."* (Musical Level 4)

> *" … The consonant categories ('s' and 'ch': lots of harmonics, jump to higher pitches; 't' and 'c': very clear; 'p': with a slight « appogiatura ») …. were much less reliable [than the vowels] (« ticket » and « piquet » one after the other were easily identifiable) …"* (Musical Level 5)

> *"[I used the] relative frequency (Hz) of certain vowels (o-e-a-i in increasing order). For the consonants, I paid especially close attention to the slides/fixed notes: in my opinion, /k/ is less of a slide than /b/ is. I used the word "kilo" as a landmark, because the attack allowed me to find the pitches for the other vowels…."* (Musical Level 6)

Such responses underline a certain awareness of phonological features in both vowels and consonants. These features are described in detail (and rather accurately) according to acoustic differences, underlining an ability to describe sounds coherent with our observations

[26] These excerpts were translated from French to English

in A.4.1. Nonetheless, this awareness, especially of vowels, is also present in participants who mentioned having more difficulties with the task, which included participants with a high level of musical experience as well as those without. They provided a less detailed description of their method:

> *"It was sometimes very difficult. The high sounds were, to me, the vowel 'i'…."* (Musical Level 4)

> *"The high pitches make me think of the sounds of 'i', 'é' and 'u' and the low pitches of 'a', 'o', 'eu'…."* (Musical Level 6)

"I tried to separate the sounds according to syllable and to associate high sounds to the "i" and low sounds to the "a"… (Musical Level 0)

The use of the phrases "to me" ("*pour moi*") or "I tried", underlines the participants' uncertainty of these descriptive traits, even though they associated pitches with vowels.

A.5.1.3 Discussion

These comments bring to light two main factors to consider in the speech perception process. The first is the role of pitch, especially associated with the vowels. Indeed, this focus on the vowels as described by both participants who succeeded at the task and those who didn't, suggests a difference in role between vowels and consonants found within the word as well as the impact of vowel interval (see Chpt.3.3). Interestingly, the opposition between "high" and "low" vowels recalls the initial description of Silbo by Trujillo (2006), all while highlighting the continuous presence of /i/ and /o/ advantages, as shown in Chapter 5.

The second factor to consider is the syllabic segment, where comparisons between duration and syllables are often cited as categorization tools. However, the elongation of

articulation through whistled speech stretches the consonants to a duration that is unusual in modal speech. In addition, changes in pitch (notably for /s/ and /t/) can easily be decomposed into parts, giving the illusion of several syllables (see Table 3, A.3). These cues may seem counterintuitive to the phonological unit of the consonant. Finally, the cues of the consonants varied according to the position (as described in Chapter 5). These factors may explain why participants had difficulty describing the whistled consonants in the word.

These responses also underline a difference according to musical experience in one's ability to qualify and describe whistled speech, in addition to hearing finer cues (as shown in the statistical analyses proposed). Indeed, as describing musical sounds is a skill taught within the French conservatory, high-level musician participants have an abundance of musical vocabulary to rely on in addition to experience analyzing sounds that other participants may not have. Such descriptions include words such as "glissando" and "appoggiatura", various mentions of vowel intervals (in terms of musical intervals), and various articulatory and acoustic aspects of consonants. Thus, these comments underline a difference in perception according to musical experience due perhaps not only to improvements in audition or pitch-related knowledge but also to the ability to organize and describe new sounds according to specific (and re-occurring) cues.

A.5.2 Consonant variability in the whistled word

Table 5:

Whistled word description

	Word	IPA form //	Target vowels //		Target cons //			Word	IPA form //	Target vowels //		Target cons //	
			V1	V2	C1	C2				V1	V2	C1	C2
1	Bateau	bato	a	o		t	2	Béquille	bekij	e	i		k
3	Cassis	kasis	a	i	k	s	4	Cocher	koʃe	o	e	k	
5	Copie	kopi	o	i	k	p	6	Chameau	ʃamo	a	o		
7	Dépôt	depo	e	o		p	8	Finale	final	i	a		
9	Fossé	fose	o	e		s	10	Kilo	Kilo	i	o	k	
11	Mégot	mego	e	o			12	Peril	peʁil	e	i	p	
13	Passé	pase	a	e	p	s	14	Petard	petaʁ	e	a	p	t
15	Piquet	pike	a	e	p	k	16	Police	polis	o	i	p	
17	Sachet	saʃe	a	e	s		18	Sauna	sona	o	a	s	
19	Sirop	siʁo	i	o	s		20	Soda	soda	o	a	s	
21	Tapis	tapi	a	i	t	p	22	Têtard	tetaʁ	e	a	t	t
23	Ticket	tike	i	e	t	k	24	Tisane	tizan	e	a	t	

Table 6:

Characterization of whistled consonant coarticulation within words

	C1			C2			C3	
Word	C1 []		Acute (Rise/Descent)	C2 []	Acute (Rise/ Descent)	Interrupted	C3 []	Acute (Rise/Descent)
bato	b			t	Yes	Yes		
bekij				k		Yes	j	No
kasis		→ a	slight descent	s	Yes		s	slight rise
koʃe	k	→ o		ʃ	Yes			
kopi				p → i	slight rise	Yes		
kilo				l → o	slight rise/descent			
depo	d	→ e	descent	p → o	slight descent	Yes		
final	f	→ i	rise	n → a	descent		l→ ə	rise
fose				s	Yes			
ʃamo	ʃ		descent	m → o	very slight descent	Yes		
mego	m			g		Yes		
peʁil	p		slight rise	ʁ → i	slight descent			slight rise
pase				s	Yes			
petaʁ		→ e		t	Yes	Yes	ʁ	slight rise
pike		→ i		k		Yes		
polis		→ l		l → i	slight rise		s	rise
saʃe	s		descent	ʃ	Yes			
sona				n	Yes			
siʁo				ʁ → o	slight descent			
soda				d	Yes	Yes		
tapi	t		descent	p	slight rise	Yes		
tetaʁ				t	Yes	Yes	ʁ	very slight rise
tike				k		Yes		
tizan			slight descent	z	descent		n	rise