INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Categorization of Whistled Consonants by French Speakers

Anaïs Tran Ngoc[1], *Julien Meyer*[2], *Fanny Meunier*[1]

[1]Université Côte d'Azur, CNRS, BCL, France
[2]Université Grenoble Alpes, CNRS, GIPSA-Lab, Grenoble, France
anais.tran-ngoc@etu.univ-cotedazur.fr, julien.meyer@gipsa-lab.fr, fanny.meunier@unice.fr

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.

Index Terms: consonant categorization, whistled speech, whistled languages

1. 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 [1, 2]. 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 [3].

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 [2, 4]. 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 [5]. More recently, Meyer et al 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 [6]. 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 [4, 7]. 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., where perceptual learning generalized to untrained word stimuli is observed for noise-vocoded speech [8]. 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/) 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 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/) [5]. 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 [2], 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

http://dx.doi.org/10.21437/Interspeech.2020-2683

the reverse [2]. 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 [2]) and the listening distance.

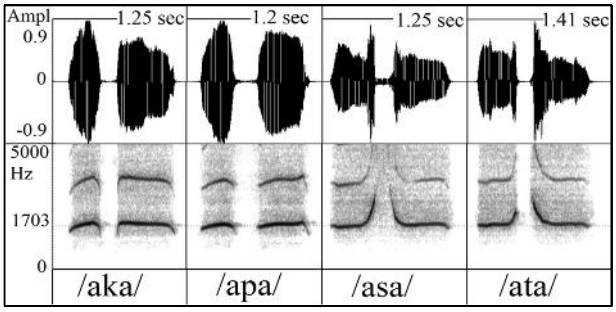

Figure 1: *Spectrogram and signal of VCV forms.*

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/ [9] 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 [9], 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).

2. Experiment 1

2.1. Method

2.1.1. Stimuli

We chose to test four distinct consonants of spoken Spanish, that have either identical or easily learned pronunciation differences in Spanish and French [10]: 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 [2]. 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 the second author.

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) [2].

2.1.2. 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, 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.

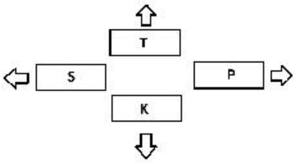

Figure 2: *Consonant/Arrow key attribution.*

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 a 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 one 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.

2.1.3. 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.

2.2. 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”.

2.3. 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<0.02). 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<0.05). These significant differences suggest a clearer recognition of “t” compared to the other consonants (Figure 3).

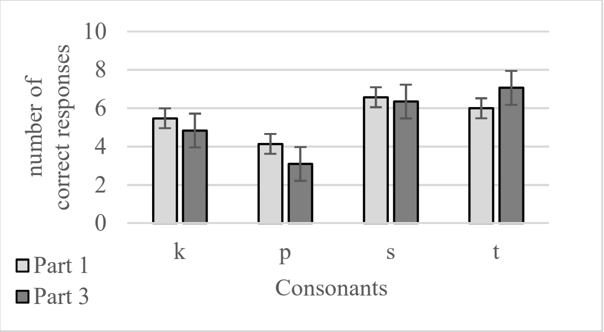

Figure 3: *Average correct responses obtained per consonant and participant in parts 1 and 3 of Experiment 1*

3. 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.

3.1. Method

3.1.1. 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 [11] 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 [12]. All participants heard the stimuli through Senheiser HD 200 Pro or Senheiser 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).

3.1.2. 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.

3.2. 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".

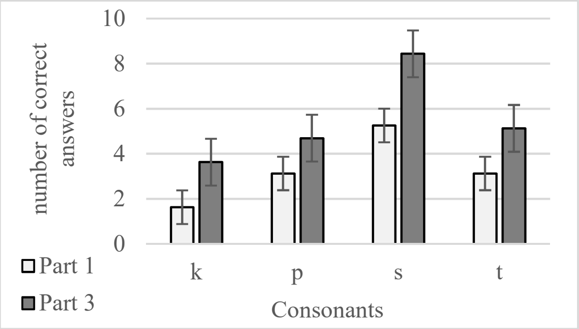

Figure 4: *Average correct responses obtained per consonant in Parts 1 and 3 of Experiment 2*

3.3. 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").

3.3.1. 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<0.001). 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.

4. 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 [1], as well as Rialland, where Silbo whistlers showed consonant preferences [5]. 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") [4]. 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) which 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?

5. 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.

6. Acknowledgements

We thank the whistler David for the consonant recordings, Jonathan Parente for his help with data collection in Experiment 2, and the participants who volunteered their time.

7. References

[1] J. Meyer, *Description Typologique et Intelligibilité des langues sifflées, approche linguistique et bioacoustique*, Thesis, 2005.

[2] J. Meyer, *Whistled Languages, A Worldwide Inquiry on Human Whistled Speech*, Springer, 2015.

[3] N. Blanco, J. Meyer, M. Hoen, F. Meunier, "Phoneme resistance and Phoneme Confusion in Noise : Impact of Dyslexia", *Interspeech*, pp.2290-2294, 2018.

[4] R. Busnel, & A. Classe, *Whistled languages*. Berlin Heideleberg: Springer-Verlag, 1976.

[5] A. Rialland, "Phonological and phonetic aspects of whistled languages," *Phonology*, Cambridge University Press (CUP), vol. 22, no. 2, pp. 237-271, 2005.

[6] J. Meyer, L. Dentel, S. Gerber, and R. Ridouane, "A Perceptual Study of CV Syllables in both Spoken and Whistled Speech: a Tashlhiyt Berber Perspective", *Interspeech*, pp. 2295-2299, 2019.

[7] J. Meyer, L. Dentel, and F. Meunier, "Categorization of Natural Spanish Whistled Vowels by Naive Listeners of Different Language Background", *Frontiers in Psychology*, vol. 8, no. 25, 2017.

[8] A. Hervais-Adelman, M. H. Dabic, I.S. Johnsrude, R.P. Carlyon, "Perceptual learning of noise vocoded words: Effects of feedback and lexicality", *Journal of Experimental Psychology: Human Perception and Performance,* vol. 34, no. 2, pp. 460–474, 2008.

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

[10] J. Molina Mejia, *Diagnostique et Correction des Erreurs de Prononciation en FLE des apprenants Hispanophones*, Mémoire M1, direction de Dominique Abry, Grenoble : Université de Stendhal, 2007.

[11] A. D. Gotzen, N. Bernardini and D. Arfib, "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, 2000.

[12] A. Tran Ngoc, F. Meunier, and J. Meyer, "Whistled vowel identification by French listeners," in *INTERSPEECH 2020 – 21th Annual Conference of the International Speech Communication Association, September 14-18, Shanghai, China, Proceedings*, 2020, pp.

[13] A. Tran Ngoc, J. Meyer, and F. Meunier, "Categorization of Whistled Consonants by Naive French Speakers," in *INTERSPEECH 2020 – 21th Annual Conference of the International Speech Communication Association, September 14-18, Shanghai, China, Proceedings*, 2020, pp.