INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

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

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

[1]Université Nice Côte d'Azur, CNRS, BCL, Nice, France
[2]Université Grenoble Alpes, CNRS, GIPSA-Lab, Grenoble, France

anais.tran-ngoc@etu.univ-cotedazur.fr, fanny.meunier@unice.fr, julien.meyer@cnrs.fr

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.

Index Terms: speech perception, whistled speech, word perception, vowels

1. 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 [1, 2]. 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 overlaping with /o/ and sometimes with /a/ [3,4,5,6]. Whistled consonants modulate these pitches through their corresponding spoken articulation. This can cause for rapid pitch changes (e.g. /s/ and /t/, Fig.1) or stops at stable pitches (e.g. for /k/ and /p/, Fig.1) [5, 6].

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 [6,7]. 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 [6].

So far, whistled word identification with naive listeners has not been studied. However, several experiments conducted on whistled speech perception showed that naive French listeners' perception of whistled phonemes is well over chance [9, 10, 11]. 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 [12, 13, 14]. 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 [12]. In general, the contribution of vowels to word recognition is mediated by context. For example, Fogerty & Humes 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 [15]. The relationship between vowels is also found to be a contributing factor for word recognition in specific cases, such as in CVCV words [16].

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) [16]. Concerning (i), the whistled vowel recognition rates obtained in previous studies [9,10,11] 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 [17]. Thus, we hypothesize that participants may have more facility with words containing larger inter-vowel intervals, particularly when

10.21437/Interspeech.2023-1967

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 [9, 10, 11]. 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.

2. Experiment

2.1. Method

2.1.1. 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).

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

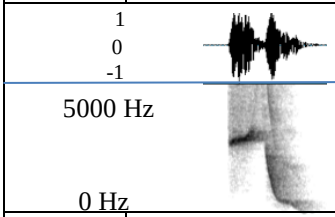
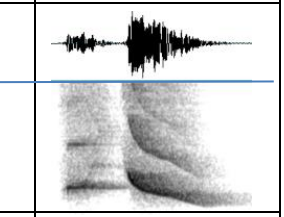
Word	/kilo/	/bato/
		
Duration	715,25 (SD=34)	918,8 (SD = 32)

Figure 1: *Signal and spectrogram of /kilo/ and /bato/*

2.1.2. 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 [6]. 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 significiant (Pearson's correlation r(22)=.37). The variability in duration between words is similar between the whistled words and the modal spoken words, though slightly lower for whistled speech. 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, 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 [6, 9]; 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 in other studies [9, 17] 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 (with 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 [1,5,6].

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 [9]). We note that this stable relative distance may compensate for the variability of the pitches.

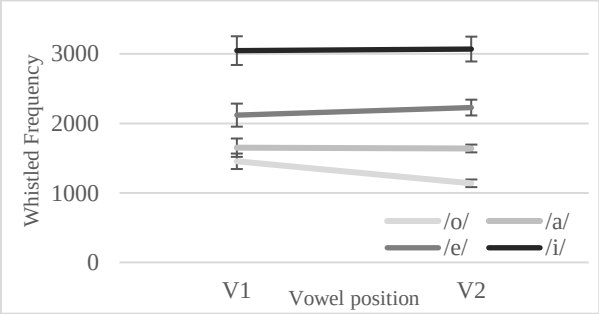


Figure 2: *Vowel frequency comparison between V1 and V2 positions in the words tested here*

2.1.3. 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 a 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.

2.1.4. 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.

2.1.5. Participants

Nineteen participants were included in this experiment 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.

2.2. Results

2.2.1. 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/).

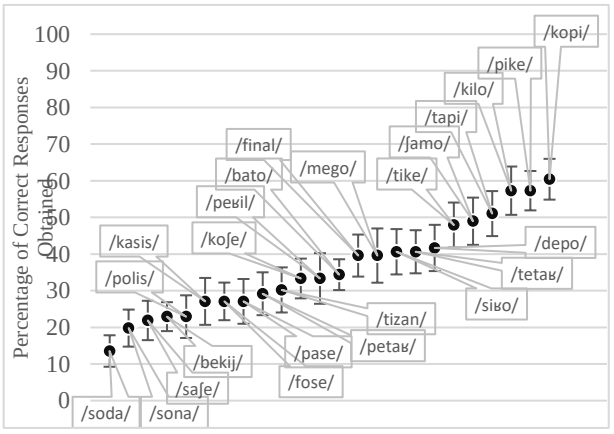


Figure 3: *Percentage of Correct responses obtained per word with standard error shown*

2.2.2 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.

2.2.3 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 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/ (ps<.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/ (ps<.001), and /o/ > /a/ and /o/ > /e/ (ps<.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).

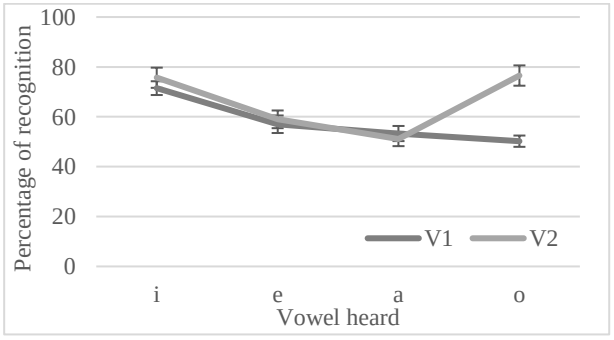


Figure 4: *Effect of Vowel position on Vowel recognition*

2.2.4 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.

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 found 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 [6, 8, 10, 11]. 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 between played and answered words, 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 (they are notably more affected by vowel position). The effect of vowel position that we found first suggests an impact of context on vowel contributions to disyllabic word recognition (in line with Delle Luche et al., [16] for spoken words). In furthering our analysis of the vowel played and its position, we found that the vowel position affects categorization hierarchy. In V1, words with /i/ are recognized better than words with any of the other vowels. However, in V2, this advantage is shared with words containing /o/, giving for this position the same vowel hierarchy as found in previous studies with vowels presented alone (/i>o>a=e/) [9, 17]. This difference between positions is very probably due to the proximity between the frequencies of /o/ and /a/ in V1. Another possible influence might be the better stability of V2 extracted values due to longer durations in the second syllable.

Finally, a more detailed exploration of the vocalic context in disyllabic words showed that larger vowel frequency differences coined better whistled word recognition for naïve listeners (we found an effect of relative vowel frequency interval, where 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 also mention this importance of relative frequencies [5], which generally contributes to increase vowel recognition [17].

4. Conclusions

In conclusion, naive listeners recognize whistled words of the test correctly - 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.

5. Acknowledgements

We would like to thank David for the whistled word recordings, and to the participants for volunteering their time.

6. References

[1] R.G. Busnel and A. Classe, *Whistled Languages*. Berlin: Springer-Verlag, 1976.

[2] J. Meyer, "Environmental and linguistic typology of whistled languages". *Annual Review of Linguistics*, 7, pp. 493-510, 2021.

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

[4] A. Classe, "Phonetics of the Silbo Gomero". *Archivum Linguisticum 9*, pp. 44–61, 1956.

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

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

[7] A. Moles, "Etude sociolinguistique de la langue sifflée de Kusköy". *Revue de Phonétique Appliquée*, 14/15, pp. 78-118, 1970.

[8] R-G. Busnel, "Recherches expérimentales sur la langue sifflée de Kusköy". *Revue de Phonétique Appliquée*, 14/15, pp. 41-57, 1970.

[9] A. Tran Ngoc, J. Meyer, F. Meunier, "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. 1605-1609

[10] A. Tran Ngoc, J. Meyer, 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. 1600-1604

[11] A. Tran Ngoc, F. Meunier, J. Meyer "Testing Perceptual Flexibility in Speech through the Categorization of Whistled Spanish Consonants by French Speakers", *JASA Express Letters* 2, 095201, 2022.

[12] J. Meyer, L. Dentel, F. Meunier, "Speech Recognition in Natural Background Noise". *PLoS ONE*, 8(11): e79279, 2013.

[13] JR. Benki, "Analysis of English Nonsense Syllable Recognition in Noise". *Phonetica* 60 pp. 129–157, 2003.

[14] L. Varnet, J. Meyer, M. Hoen, F. Meunier, "Phoneme resistance during speech-in-speech comprehension", *Proceedings of Interspeech Portland, USA*, 2012.

[15] D., Fogerty & LE., Humes, "Perceptual contributions to monosyllabic word intelligibility: segmental, lexical, and noise replacement factors". *J Acoust Soc Am* 128, pp. 3114–3125, 2010.

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

[17] J. Meyer, "Acoustic Strategy and Typology of Whistled Languages; Phonetic Comparison and Perceptual Cues of Whistled Vowels". *Journal of the International Phonetic Association. Cambridge University Press*, vol. 38, no.1, pp. 69-94, 2008