Original Article

Language *and* Speech

Statistical Speech Segmentation in Tone Languages: The Role of Lexical Tones

Language *and* Speech
1–13

DOI: 10.1177/0023830917706529
journals.sagepub.com/home/las
SAGE

David M. Gómez
Institute for Educational Sciences, Universidad de O'Higgins, Chile; Center for Advanced Research in Education (CIAE), Universidad de Chile, Chile

Peggy Mok
Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong

Mikhail Ordin
Basque Centre on Cognition, Brain, and Language (BCBL), Spain; Basque Foundation for Science (IKERBASQUE), Spain

Jacques Mehler
Language, Cognition, and Development Lab, International School for Advanced Studies (SISSA), Italy

Marina Nespor
Language, Cognition, and Development Lab, International School for Advanced Studies (SISSA), Italy

Abstract
Research has demonstrated distinct roles for consonants and vowels in speech processing. For example, consonants have been shown to support lexical processes, such as the segmentation of speech based on transitional probabilities (TPs), more effectively than vowels. Theory and data so far, however, have considered only non-tone languages, that is to say, languages that lack contrastive lexical tones. In the present work, we provide a first investigation of the role of consonants and vowels in statistical speech segmentation by native speakers of Cantonese, as well as assessing how tones modulate the processing of vowels. Results show that Cantonese speakers are unable to use statistical cues carried by consonants for segmentation, but they can use cues carried by vowels. This difference becomes more evident when considering tone-bearing vowels. Additional data from speakers of Russian and Mandarin suggest that the ability of Cantonese speakers to segment streams with statistical cues carried by tone-bearing vowels extends to other tone languages, but is much reduced in speakers of non-tone languages.

Corresponding author:
David Maximiliano Gómez, Instituto de Ciencias de la Educación (ICEd), Universidad de O'Higgins, Avenida Libertador Bernardo O'Higgins 611, 2841959 Rancagua, Chile.
Email: david.gomez@uoh.cl

Keywords
Lexical tone, tone language, transitional probability, speech segmentation, consonants and vowels

1. Introduction

Spoken languages use consonants and vowels to form words. They also use pitch modulations, which can occur at the level of the phrase (intonation) or of the word (lexical tone). Whereas variations in intonation are used to convey phrasal prosodic structure or emotion, tonal variation is also used by some languages to signal lexical contrasts. Cantonese is an example of these *tone languages*, where the syllable [ji] means either doctor (醫), chair (椅), idea (意), child (兒), ear (耳), or two (二), depending on the tone with which it is produced. Although the processing of lexical tone is becoming a prolific field of inquiry (e.g., Singh, Goh, & Wewalaarachchi, 2015; Singh, Hui, Chan, & Golinkoff, 2014), our knowledge of tone languages and how tone modulates speech perception is still limited because of a lack of data from tone languages.

Of particular interest to us is how the presence of lexical tone in a language affects speech-processing biases that have been widely documented in non-tone languages. For instance, tone languages exhibit reversed effects in the processing of consonants and vowels when compared to non-tone languages: it is a well-known fact that when speakers of Dutch and Spanish are presented with a pseudoword, such as *kebra*, and asked to convert it into a word by changing a single phoneme, they prefer to substitute a vowel (e.g., producing *cobra*) rather than a consonant (e.g., producing *zebra*) (Cutler, Sebastián-Gallés, Soler-Vilageliu, & Van Ooijen, 2000). This bias, however, is reversed in speakers of Mandarin, as shown by Wiener and Turnbull (2016). In the present study, we explore a similar case of processing bias regarding consonants and vowels that has been investigated in several non-tone languages but not, so far, in tone languages, namely that consonants are favored for the computation of transitional probabilities (TPs) in speech segmentation.

1.1 Consonants, vowels, and their contribution to speech segmentation

On the basis of cross-linguistic comparisons, Nespor, Peña, and Mehler (2003) proposed that consonants and vowels have distinct roles in language: Consonants would support lexical processing, whereas vowels would convey structural properties. Empirical support for the so-called CV hypothesis was put forward by Bonatti, Peña, Nespor, and Mehler (2005), where an artificial language segmentation task based on the computation of TPs was used to tap lexical processes in French adults. In these tasks, participants are exposed to a speech stream without pauses or other prosodic cues signaling word boundaries, leaving TPs between consecutive units—typically syllables—as the only cues to word structure. TP-based speech segmentation may be regarded as a lexical process used to extract words from continuous speech, hence the CV hypothesis predicts that performance in this task benefits more from information carried by consonants than from information carried by vowels. Bonatti et al. (2005) used experimental materials specifically designed to dissociate TPs between consecutive syllables, consonants, and vowels and evaluate the capacity of consonants and vowels to support successful segmentation. As predicted by the CV hypothesis, they observed that participants succeeded in using statistical information carried by consonants, but were unable to do the same when statistical information was carried by vowels. Mehler, Peña, Nespor, and Bonatti (2006) presented further evidence of French listeners' preference to rely on TPs carried by consonants rather than on those carried by vowels for statistical

segmentation (see also Toro, Nespor, Mehler, & Bonatti, 2008, who tested Italian speakers). These researchers suggested that the functional specializations of consonants and vowels posited by the CV hypothesis reflect a universal psycholinguistic bias.

1.2 The present study

In this article, we test whether the CV hypothesis can be extended to a language with lexical tones. Our first research question was whether speakers of a tone language are sensitive to TPs carried by consonants in the same way as speakers of non-tone languages. We explored this question by presenting native speakers of Cantonese with flat-tone speech streams carrying TP cues either only on consonants or only on vowels. Our second question involved the effect of tonal variation in the speech streams and, in particular, we investigated whether the computation of TPs on vowels and tones coupled together would be improved with respect to vowels alone. We then further explored the generalizability of Cantonese speakers' data by testing speakers of Mandarin and Russian (a tone and a non-tone language, respectively) with two conditions that we observed to be critical. To our knowledge, this is the first study using the TP-based speech segmentation paradigm to assess the CV hypothesis with speakers of tone languages.

1.3 Tones in Hong Kong Cantonese

Experiments 1 and 2 studied the ability of Cantonese speakers to extract TPs from speech streams. Cantonese is a Sino-Tibetan language with 19 consonants and 8 vowels (e.g., Cheung, 1972; Yue-Hashimoto, 1972). In addition, Cantonese has a rich tonal system comprising six contrastive lexical tones (T1 to T6) based on pitch patterns alone on open syllables or syllables with nasal endings (Bauer & Benedict, 1997; Fok-Chan, 1974)[1]. These tones can be schematically described in terms of their starting and ending pitch levels. For instance, if pitch levels 1 and 5 denote the lowest and highest fundamental frequency (F0) levels for a given speaker, T1 can be described as [55] or a high-level tone (Figure 1; see also Chao, 1930). Cantonese has three level tones (T1, T3, T6), two rising tones (T2, T5), and one falling tone (T4). In the following, Experiment 1 used only the low-level tone, T6. Experiment 2 used, instead, tones T1, T2, T3, and T4.

2. Experiment 1

We investigated our first research question by presenting Cantonese speakers from Hong Kong with artificial speech streams where statistical cues were carried either only by consonants (C group) or only by vowels (V group). To make our results comparable with previous data from non-tone languages (see Table 1 for a short summary of relevant data), we used only syllables produced with a low-level tone (T6; see Methods). To avoid any confusion due to syllable frequency in Cantonese, all syllables used in the streams were unattested in the language (bearing this specific tone).

2.1 Methods

2.1.1 Participants. The participants in Experiment 1 were undergraduate students of linguistics at The Chinese University of Hong Kong, who received course credit for their participation. They were Cantonese–English[2] bilinguals but reported Cantonese to be their dominant language, as well as having no auditory problems. The C group comprised 21 adults (16 women, mean age 20.2 ± 2.7 years, age range 18–30), whereas the V group comprised 21 adults (17 women, mean age 20.0 ±

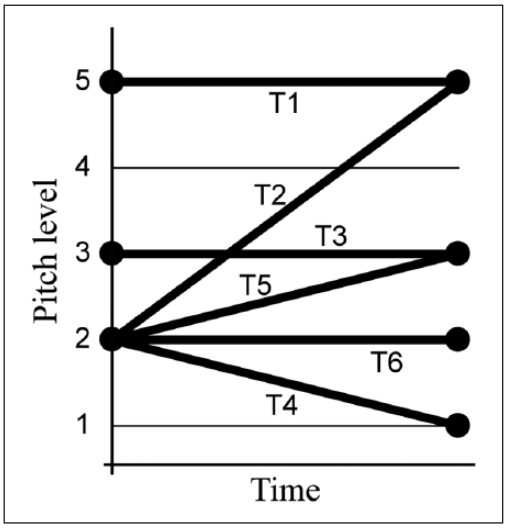

Figure 1. Schematic depiction of the six Cantonese tones. Pitch levels from 1 (lowest) to 5 (highest) are relative to a given speaker's F0 range (for instance, T2 is a rising tone that starts at pitch level 2 and finishes at level 5).

Table 1. Average scores for Experiments 1–4 *$p < .05$ **$p < .01$ ***$p < .001$.

Experiment	Language	Tone	Group	% correct (*SD*)	Effect size (Cohen's d)
1	Cantonese	Flat	C	51.4% (13.7%)	0.10
	Cantonese	Flat	V	54.6% (10.0%)*	0.46
2	Cantonese	Varying	V+T	68.4% (14.5%)***	1.27
3	Russian	Flat	C	61.8% (10.1%)***	1.17
	Russian	Varying	V+T	55.4% (8.1%)**	0.66
4	Mandarin	Flat	C	59.0% (9.6%)***	0.93
	Mandarin	Varying	V+T	70.0 % (8.3%)***	2.41

2.2, age range 18–26). Two additional participants were tested but not included because they declared English to be their dominant language.

The Ethics Committee of the International School for Advanced Studies (SISSA) in Trieste, Italy, approved the protocols of all the experiments of this study.

2.1.2 Materials. We synthesized two artificial speech streams as follows: We built a set of 16 CV syllables obtained by combining the consonants b, g, d, and h ([p], [k], [t], and [h] respectively)[3] and the vowels e, i, u, and y ([ɛ], [i], [u], and [y] respectively). Each syllable in the set was spoken by a phonetician—a native speaker of Hong Kong Cantonese—with a low-level tone (T6, see Figure 1).

Recorded syllables were then resynthesized with Praat (Boersma, 2001) in order to standardize them to a length of 450 ms and a flat pitch contour at 148 Hz, and later concatenated in pseudorandom order to make speech streams of an approximate total duration of 860 s. Splicing was done with SoX Sound eXchange version 14.4.0 (http://sox.sourceforge.net/), using a quarter-cosine wave as fade envelope, and excess and leeway parameters of 5 ms and 10 ms, respectively.

The order in which syllables were concatenated was chosen so as to keep TPs between syllables uninformative, and to convey useful statistical information only between either consecutive consonants (for the C group) or consecutive vowels (for the V group). More specifically, speech streams presented to the C group were built so that: (a) TPs between adjacent syllables were always 0.5, for instance Prob(ge → du) = 0.5 and Prob(ge → dy) = 0.5; (b) TPs between consecutive vowels were always 0.5, for instance Prob(y → i) = 0.5 and Prob(y → e) = 0.5; and (c) TPs between consecutive consonants were either 0.75 or 0.25, for instance Prob(h → b) = 0.75 and Prob(h → g) = 0.25. That is to say, TPs were flat and uninformative when computed either at the syllable level or between consecutive vowels, whereas TPs between consecutive consonants were useful for segmentation (e.g., *hebi* had a greater consonant-TP than *hegi*, 0.75 and 0.25 respectively). An analogous procedure yielded the speech stream used for the V group, where TPs between consecutive vowels were informative and both TPs between syllables and TPs between consecutive consonants were uninformative. In the Appendix we provide an example of how to build such speech streams.

Segmentation was tested by means of a two-alternative forced choice task presented at the end of the listening phase. This test comprised 24 items contrasting bisyllables with either a consonant-TP of 0.75 (for the C group) or a vowel-TP of 0.75 (for the V group), against bisyllables with the relevant TP of 0.25 or lower. To make sure that we assessed TP extraction and not just memory for specific chunks of syllables, we used only bisyllables that never appeared in the speech streams for the test.

2.1.3 Procedure. Each speech stream was divided into two parts of about 430 s. In order to avoid any cues in addition to statistical information, the first and last 10 seconds of each part were faded in and out respectively, using a quarter-sine wave envelope. Instructions were presented on the computer screen in both Cantonese and English. The two streams were presented separated by a self-paced pause. After the listening phase, participants were instructed about the test. In the test, they listened to two of the test bisyllables described above one after the other, and selected the one that belonged to the artificial language they had just heard. They did so by using the keys 1 and 2, standing for the first and the second word, respectively. Keys associated to correct responses were counterbalanced. The order of presentation of test items was pseudorandomized with the constraint that no more than two items in a row could share the same response key.

2.2 Results and discussion

Table 1 presents the main results for all experiments. Participants in the C group were unable to segment the speech streams, $M = 51.4\%$, $t(20) < 1$, *n.s.*, in contrast to previous reports of test participants speaking non-tone languages (see Table 1). Perhaps even more striking is the outcome of participants in the V group, who performed significantly above chance in the segmentation test, $M = 54.6\%$, $t(20) = 2.09$, $p = .05$. This result increases the contrast with speakers of non-tone languages (Bonatti et al., 2005; cf. Newport & Aslin, 2004). Nonetheless, it is important to note that the magnitude of this effect was very small (less than 5% above chance level) with respect to previous statistical segmentation studies (see Table 2 for reference values). In line with this, the performance of the C and V groups was statistically undistinguishable, $t(40) < 1$, *n.s.*.

Cantonese speakers' performance in the statistical segmentation task departed significantly from previous data obtained from speakers of non-tone languages, as evidenced by the fact that participants were able to make use of TPs carried by vowels but not of those carried by consonants. This cross-linguistic difference was unpredicted by Nespor et al.'s (2003) CV hypothesis, which was based on evidence exclusively from non-tone languages. These authors argued that the consonantal bias observed in non-tone languages stems from the categoricity of consonant

Table 2. Average scores for some statistical segmentation studies that have addressed the contribution of consonants and vowels separately.

Study	Language	Relevant unit	% correct (*SD*)	Effect size (Cohen's d)
Newport & Aslin (2004) (Exp. 2)	English	C	78.4% (18.9%)	1.50
Newport & Aslin (2004) (Exp. 3)	English	V	79.1% (19.2%)	1.52
Bonatti et al. (2005) (Exp. 1)	French	C	87.7% (10.6%)	3.56
Bonatti et al. (2005) (Exp. 2a)	French	V	54.2% (11.7%)	0.36
Bonatti et al. (2005) (Exp. 2b)	French	V	54.0% (11.3%)	0.35
Bonatti et al. (2005) (Exp. 2c)	French	V	70.5% (27.8%)	0.73
Toro et al. (2008) (Exp. 1)	Italian	C	63.3% (8.7%)	1.53
Toro et al. (2008) (Exp. 2)	Italian	V	51.6% (14.1%)	0.11

perception. To the best of our knowledge, there is no indication in the literature that Cantonese speakers' categorical perception of consonants might be reduced with respect to that of speakers of other languages. Although the performance of the V group was low compared to other segmentation studies, our results so far are consistent with the hypothesis that the presence of lexical tones in a language alters several patterns of processing previously thought to hold universally (cf. Wiener & Turnbull, 2016).

3. Experiment 2

Our second experiment explored more directly the role of lexical tones in segmentation by introducing tonal variation in our speech streams. More specifically, we investigated whether Cantonese speakers' performance is enhanced if TPs are carried by vowels and tones together, rather than just vowels alone. To do this, we presented a new group of Cantonese speakers with speech streams where vowels were paired with tones in a one-to-one manner (see Methods).

3.1 Methods

3.1.1 Participants. The participants in Experiment 2 were also undergraduate students of linguistics at The Chinese University of Hong Kong and received course credit for their participation. As in Experiment 1, they were Cantonese–English bilinguals who reported Cantonese to be their dominant language, as well as having no auditory problems. We tested a group composed of 19 adults (14 women, mean age 19.4 ± 1.6, age range 18–24). One additional participant was not included in the analysis because his/her score in the test was more than 2.5 standard deviations below average.

3.1.2 Materials. Speech streams were built in a similar manner to those of the V group in Experiment 1, but this time each syllable was recorded with a specific tone. We paired all e vowels with T3 (mid-level tone), all i vowels with T1 (high-level tone), all u vowels with T4 (falling tone), and all y vowels with T2 (high-rising tone). The recording was done by the same phonetician who recorded the material for Experiment 1. As before, we made sure that none of the resulting syllables were attested in Cantonese. Syllables were then digitally modified to have a length of 450 ms and the pitch contours presented in Figure 2, and later concatenated into speech streams as in the previous experiment.

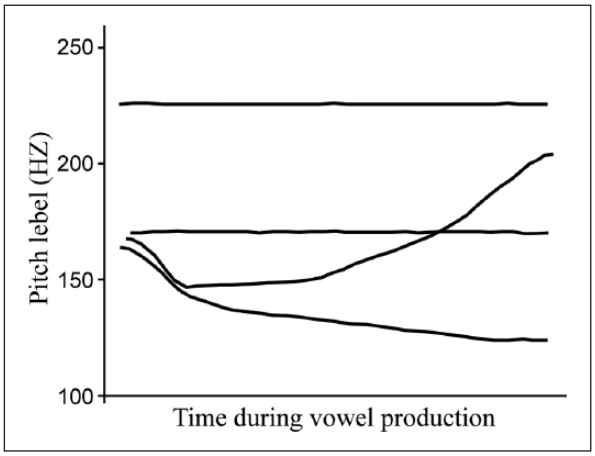

Figure 2. Pitch contours for the tones used in the streams with tonal variation in Experiments 2, 3, and 4.

3.1.3 Procedure. The procedure was similar to that used for Experiment 1.

3.2 Results and discussion

The participants in Experiment 2 were successful in using TPs carried by vowels and tones together, $M = 68.4\%$, $t(18) = 5.55$, $p < .0001$. This result was also significantly higher than that of the V group in Experiment 1, $t(38) = 3.55$, $p = .001$, suggesting that tones play a relevant role in Cantonese speakers' computation of TPs on vowels. These data provided further confirmation of the cross-linguistic differences with non-tone languages observed in Experiment 1, indicating that the CV hypothesis as proposed in its original form (Nespor et al., 2003) needs to be reevaluated.

It is important to note that our decision of pairing vowels and tones in Experiment 2 makes it difficult to isolate the effect of tones in our results. However, we observed clear evidence that TP extraction from vowels and tones together is significantly superior to that from vowels alone. This enhancement due to tonal variation may be due to an ability of Cantonese speakers to specifically process statistical cues on lexical tones, although alternative accounts are possible. Saffran, Johnson, Aslin, and Newport (1999) showed that adults can segment sequences of pure tones on the basis of TPs, opening the question of whether the improvement in segmentation is due to the accumulation of statistical cues from vowels and pitch. From this point of view, other acoustic cues that co-varied with the vowels could have improved performance in a similar manner. To achieve a more complete understanding of the effect of tones in statistical segmentation, we conducted a third experiment with speakers of a non-tone language.

4. Experiment 3

Experiments 1 and 2 revealed that speakers of Cantonese were unable to use statistical information carried by consonants, in sharp contrast to statistical information carried by vowels and tones together. These results raise two further questions. A first methodological question is whether speakers of a non-tone language would succeed in segmenting the speech streams where TPs are carried by consonants. This issue is important to assert the validity of our materials and comparability with previous research. The second question relates to the good segmentation performance

by Cantonese speakers when using statistical cues carried by tone-bearing vowels, and if this outcome is specific to speakers of tone languages. To investigate these issues we recruited native speakers of Russian, a non-tone language whose speakers can distinguish all the syllables in our materials (even if they might perceive different segments than speakers of Cantonese). Russian speakers were randomly assigned to one of two groups, one that listened to the same materials provided for the C group in Experiment 1 (flat intonation and TPs carried by consonants), and a second one that listened to the same materials provided for Experiment 2 (TPs carried by tone-bearing vowels).

4.1 Methods

4.1.1 Participants. Participants were undergraduate students or young professionals in areas related to mathematics and computer science from Lipetsk, Russia, and received a monetary reward for their participation. All of them were monolingual speakers of Russian and reported no auditory problems. The C group was comprised of 24 adults (14 women, mean age 23.7 ± 3.6 years, age range 18–30), whereas the V+T group was comprised of 24 adults (12 women, mean age 24.4 ± 4.5, age range 18–32). Two additional participants were tested but not included because they failed to comply with the instructions in the test.

4.1.2 Materials. For this experiment, we used the speech streams provided for the C group in Experiment 1 and those provided for Experiment 2.

4.1.3 Procedure. The procedure was similar to that of Experiment 1, except that instructions to participants were presented on the computer screen in Russian.

4.2 Results and discussion

Russian speakers preferred high-TP words in both conditions, $M = 61.8\%$, $t(23) = 5.72$, $p < .0001$, V+T group: $M = 55.4\%$, $t(23) = 3.25$, $p = .003$, with a higher performance when TPs were carried by consonants than when they were carried by tone-bearing vowels, $t(46) = 2.43$, $p = .02$. These results confirm that our consonant-based speech streams are valid and replicate previous findings in similar segmentation tasks with speakers of non-tone languages (Bonatti et al., 2005; Mehler et al., 2006; Newport & Aslin, 2004; Toro et al., 2008) in agreement with the CV hypothesis, allowing us to conclude that the failure of Cantonese speakers to extract statistical cues from consonants is specific to that language group. In addition, the data from participants in the V+T group show that speakers of non-tone languages are also able to extract TPs carried by tone-bearing vowels.

We then contrasted Cantonese speakers' and Russian speakers' segmentation performance by means of a logistic regression with two factors: language (Cantonese/Russian); and statistically relevant unit (C/V+T). This analysis revealed a significant interaction between both factors ($b = -0.98$, $z = -5.45$, $p < .0001$), as Russian speakers outperformed Cantonese speakers in segmenting flat streams with TPs carried by consonants, $t(43) = 2.92$, $p = .005$ whereas Cantonese speakers outperformed Russian speakers when TPs were carried by tone-bearing vowels, $t(41) = -3.75$, $p = .0006$. This suggests that the cognitive mechanisms underlying segmentation based on TPs differ between both language groups. This pattern of results might reflect differential involvement of general auditory mechanisms (such as those allowing English speakers to segment statistically structured sequences of pure tones, see Saffran et al., 1999) and language-specific mechanisms.

5. Experiment 4

Finally, we tested the generalizability of these results to tone languages other than Cantonese. Our fourth experiment collected data from speakers of Mandarin, a tone language sharing many properties with Cantonese although the two languages are mutually unintelligible.

5.1 Methods

5.1.1 Participants. Participants were MA students of linguistics at The Chinese University of Hong Kong who received a monetary reward for their participation. They were native speakers of Mandarin and reported no auditory problems. The C group was comprised of 20 adults (all women, mean age 23.1 ± 1.5 years, age range 22–27), whereas the V+T group was comprised of 19 adults (17 women, mean age 24.0 ± 1.6, age range 22–27). One additional participant in the V+T group had a score of more than 2.5 standard deviations below average in the test and, therefore, was not included.

5.1.2 Materials. We used the same materials as in Experiment 3.

5.1.3 Procedure. The procedure was similar to that of Experiment 1, except that instructions to participants were presented on the computer screen both in Mandarin and in English.

5.2 Results and discussion

Mandarin speakers, as did Russian speakers, succeeded in segmenting both types of speech streams, C group: $M = 59.0\%$, $t(19) = 4.18$, $p = .0005$, V+T group: $M = 70.0\%$, $t(18) = 10.5$, $p < .0001$, but exhibited higher performance in extracting TPs carried by vowels and tones together than TPs carried by consonants, $t(37) = 3.82$, $p = .0005$.

We again conducted cross-linguistic comparisons by means of logistic regressions. The contrast between Mandarin and Cantonese speakers showed a significant main effect of language ($b = 0.31$, $z = 2.38$, $p = .02$), indicating that Mandarin speakers obtained higher scores than Cantonese speakers overall, as well as a significant main effect of the statistically relevant unit ($b = 0.72$, $z = 5.34$, $p < .0001$), revealing that both language groups perform better in extracting TPs carried by tone-bearing vowels than TPs carried by consonants. We observed no significant interaction between these factors ($b = -0.23$, $z = -1.22$, $p = .22$). Planned comparisons revealed that Mandarin speakers performed better than Cantonese speakers in extracting TPs carried by consonants, $t(39) = 2.04$, $p = .048$, but no difference appeared for TPs carried by tone-bearing vowels, $t(36) < 1$, *n.s*. These findings reveal commonalities and specificities of tone languages regarding statistical segmentation: Both language groups performed significantly better in segmenting streams based on tone-bearing vowels than those based on consonants, but only Mandarin speakers succeeded in using the latter type of cues, underlining the importance of studying these languages more extensively. It is possible that Mandarin speakers performed better because there are more consonants in Mandarin than in Cantonese (24 vs. 19) (Lee & Zee, 2003; Zee, 1999), so they may be more accustomed to extracting TPs carried by consonants.

The contrast between Mandarin and Russian speakers, instead, showed a significant interaction ($b = -0.75$, $z = -4.09$, $p < .0001$). This effect reflects the fact that both groups segmented flat streams with TPs on consonants with similar success, $t(42) < 1$, *n.s*, but Mandarin speakers did better in segmenting tonal streams with statistical cues on tone-bearing vowels, $t(41) = 5.80$, $p < .0001$. Altogether, this suggests that Mandarin speakers process TPs carried by consonants in a similar

way to speakers of non-tone languages, whereas segmentation based on TPs carried by tone-bearing vowels appears to be achieved by engaging different cognitive processes on the part of speakers of tone and non-tone languages.

6. General discussion

Throughout this article, we have investigated the role of consonants, vowels, and lexical tones in the computation of statistical regularities present in artificial speech streams, with an emphasis on the performance of speakers of tone languages. Our data showed that Cantonese speakers failed to segment speech streams where the relevant TPs were carried by consonants alone. In contrast, they performed above chance level when TPs were carried by vowels, with a substantial improvement when vowels were paired with tones. The consonantal results stand in sharp contrast to previous evidence from non-tone languages, where test participants showed successful segmentation when consonants were statistically relevant (Bonatti et al., 2005; Mehler et al., 2006; Newport & Aslin, 2004; Toro et al., 2008). This suggests that tonal information provided strong cues for Cantonese speakers to extract the statistical cues present in the streams, beyond those signaled by vowels alone.

Two additional experiments with speakers of Mandarin—another tone language—and Russian—a non-tone language—shed further light on the generalizability of these results and the nature of the mechanisms involved. Both speakers of Mandarin and Russian succeeded in using statistical information carried by consonants, demonstrating that the ability to use this type of information is not incompatible with the presence of lexical tones in a language. Nonetheless, speakers of Mandarin and Cantonese performed similarly well in extracting TPs from tone-bearing vowels. Speakers of these two tone languages showed better segmentation of TPs carried by tone-bearing vowels than of those carried by consonants. Likewise, performance of these speakers in segmenting streams based on tone-bearing vowels was better than that of Russian speakers in the same task.

Data from Experiments 1 and 2 raise the issue of whether any acoustic cue correlated with vowels (not necessarily tones) would enhance segmentation. In this view, tones would have no special status other than improving vowels' acoustic saliency. While our data does not deal directly with this issue, they may provide a partial answer. First, a purely acoustic account should generalize similarly across languages. However, while Russian speakers extracted TPs from tone-bearing vowels significantly above chance level, their accuracy was much lower than that of speakers of Cantonese and Mandarin (55.4% vs. 68.4% and 70.0%, respectively). Second, it needs to be pointed out that tones have a complex, multidimensional acoustic realization that differs in quality (e.g., raising, falling, flat, dipping), in contrast to other dimensions such as amplitude or vowel length that differ only in quantity (e.g., higher/lower, longer/shorter). This asymmetry resembles that between consonants and vowels (Nespor et al., 2003), and deserves further research.

Altogether, our findings show that speakers of the tested tone languages differ in terms of their sensitivity to statistical cues carried by consonants, but that they both benefit to a large extent from similar TP cues carried by tone-bearing vowels. Tone-bearing vowels allow speakers of both languages to extract statistical words from artificial speech streams more successfully than speakers of a non-tone language, suggesting that the mechanisms supporting statistical segmentation in speakers of a tone language differ from those used by speakers of non-tone languages. We conjecture that speakers of non-tone languages may process statistics on tone-bearing vowels by means of general auditory mechanisms only (such as those enabling the computation of TPs between adjacent tones, Saffran et al., 1999, or nonlinguistic sounds, Gebhart, Newport, & Aslin, 2009), whereas speakers of tone languages may further activate language-specific mechanisms boosting TP computations.

It is important to point out a limitation of our experimental approach. Our materials were designed so as to provide informative statistical cues only between a single type of unit, keeping

TPs between the other units and between adjacent syllables equal to 0.5. On top of this, we used only syllables unattested in Cantonese to avoid any semantic confusion. These strong constraints limited the number of available combinations between vowels, consonants, and tones in a manner that made it extremely difficult to manipulate TPs for vowels and TPs for tones independently. As a result, the question concerning whether tones alone are sufficient carriers of statistical information is still open. Still, our data demonstrate that the presence of lexical tone in a language modulates the computation of statistical cues.

6.1 Tone languages and the CV hypothesis

Our findings constitute strong evidence that the current formulation of the CV hypothesis (Nespor et al., 2003) needs to be modified in order to include tone languages. The consonantal bias for lexical processes, one of the two aspects of the CV hypothesis, was originally proposed by Nespor et al. (2003) on the basis of the following cross-linguistic observations: (a) consonants tend to outnumber vowels; (b) in many languages, consonants tend to disharmonize within words whereas vowels tend to harmonize; (c) consonants are perceived categorically while vowels are not; and (d) sequences of consonants (but not vowels) may constitute morphological roots in some languages. None of the observations in their review included tone languages and, hence, the possibility of a functional specialization for lexical tones and their effect on the linguistic division of labor between consonants and vowels were not discussed.

To this end, it is necessary to study a broader set of tone and non-tone languages to understand the elements determining speakers' ability to use statistical cues carried by consonants and vowels. One particularly interesting case is Danish: Højen and Nazzi (2016) presented data from 20-month-old infants exposed to Danish—a language with the peculiarity that vowels outnumber consonants—in a word-learning task, showing that they tend to rely on vowels more than on consonants. If an analogous finding were demonstrated in adult speakers as well, Danish would also constitute a violation of the CV hypothesis. Another interesting case is Japanese, where pitch accent might, in principle, play a similar role to tones in Cantonese and Mandarin.

6.2 The role of lexical tones

Contrastive lexical tone is a constitutive element of many languages, but the impact of its presence on linguistic computations is not yet well understood. We have taken a first step in this direction by investigating statistical segmentation processes in Cantonese. Our data show that Cantonese speakers failed to segment streams with TPs carried by consonants, and that tones boosted Cantonese speakers' performance in segmentation based on vowels. These findings reveal important cross-linguistic differences unpredicted in the previous proposal by Nespor et al. (2003), adding to the evidence for the differences in linguistic processing between tone and non-tone languages (e.g., Wiener & Turnbull, 2016). We have also shown that the ability of vowels and tones together to sustain TP computations appears to be specific to tone languages, as it is also present in Mandarin speakers but not (or, at least, present to a lesser degree) in speakers of a non-tone language tested with the same materials. Altogether, our study contributes data to the understanding of tone languages and of how speakers of these languages process speech streams. These data highlight differences between non-tone and tone languages, including aspects that were previously thought to be universal.

Acknowledgements

The authors are grateful to Donghui Zuo for her assistance in collecting the data, and to Jean-Rémy Hochmann and anonymous reviewers for valuable comments on previous versions of this manuscript.

Funding

The research leading to these results has received funding from: the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC [grant agreement 269502, PASCAL]; the Chilean CONICYT program PIA/BASAL [grant FB0003]; the Alexander von Humboldt Foundation; and the Basque Foundation for Science (Ikerbasque).

Notes

1. Traditional Chinese phonology included three checked tones as well (i.e., short tones on syllables with final unreleased stops), thus yielding an inventory of nine tones. These checked tones, however, are considered as allotones in the six-tone system.
2. English is used extensively in Hong Kong's school system.
3. The common Romanization systems of Cantonese use b, g, and d for the voiceless plosives [p], [k], and [t] respectively. In contrast, p, k, and t stand for the aspirated voiceless plosives [p^h], [k^h], and [t^h]. All the consonants we selected for our material are voiceless and thus do not bear tonal information, restricting tonal information only to the vowels accompanying them.

References

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese Phonology*. Berlin, Germany: Mouton de Gruyter.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Bonatti, L. 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.

Chao, Y. R. (1930). A system of tone-letters. *Le Maître Phonétique*, *45*, 24–27.

Cheung, H. N. S. (1972). *Xianggang Yueyu yufa de yanjiu* [Cantonese as spoken in Hong Kong]. Shatin, Hong Kong: The Chinese University Press.

Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., & Van Ooijen, B. (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition*, *28*(5), 746–755.

Fok-Chan, Y. Y. (1974). *A Perceptual Study of Tones in Cantonese*. Shatin, Hong Kong: Hong Kong University Press.

Gebhart, A. L., Newport, E. L., & Aslin, R. N. (2009). Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychonomic Bulletin & Review*, *16*(3), 486–490.

Højen, A., & Nazzi, T. (2016). Vowel bias in Danish word-learning: Processing biases are language-specific. *Developmental Science*, *19*(1), 41–49.

Lee, W.-S., & Zee, E. (2003). Standard Chinese (Beijing). *Journal of the International Phonetic Association*, *33*(1), 109–112.

Mehler, J., Peña, M., Nespor, M., & Bonatti, L. (2006). The "soul" of language does not use statistics: Reflections on vowels and consonants. *Cortex*, *42*(6), 846–854.

Nespor, M., Peña, M., & Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e Linguaggio*, *2/2003*, 203–229.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Singh, L., Goh, H. H., & Wewalaarachchi, T. D. (2015). Spoken word recognition in early childhood: Comparative effects of vowel, consonant, and lexical tone variation. *Cognition*, *142*, 1–11.

Singh, L., Hui, T. J., Chan, C., & Golinkoff, R. M. (2014). Influences of vowel and tone variation on emergent word knowledge: A cross-linguistic investigation. *Developmental Science*, *17*(1), 94–109.

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.

Wiener, S., & Turnbull, R. (2016). Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. *Language and Speech*, *59*(1), 59–82.

Yue-Hashimoto, A. O. (1972). *Cantonese (The Yue Dialects I)*. Cambridge, UK: Cambridge University Press.

Zee, E. (1999). Chinese (Hong Kong Cantonese). In International Phonetic Association (Ed.), *Handbook of the International Phonetic Association: A Guide to the use of the International Phonetic Alphabet* (pp. 58–60). Cambridge, UK: Cambridge University Press.

Appendix

The speech streams we used had to fulfill several restrictions in order to test speakers' sensitivity to TPs carried by vowels and consonants. In Table 3, we provide an example of a transition table that was used to generate some of the speech streams where statistical cues were carried by consonants. If all transitions are presented equally often, these yield a stream where TPs between syllables are always 0.5, TPs between consecutive vowels are always 0.5, and TPs between consecutive consonants are either 0.75 or 0.25.

Table 3. Example of a possible transition table used to create a speech stream for the C group.

Syllable	May be followed by …		Syllable	May be followed by …	
Be	Gu	Gy	Ge	Du	Dy
Bi	Gu	Gy	Gi	Du	Dy
Bu	De	Gi	Gu	De	Hi
By	Di	Ge	Gy	Di	He
De	Hu	Hy	He	Bu	By
Di	Hu	Hy	Hi	Bu	By
Du	Be	Hi	Hu	Be	Gi
Dy	He	Bi	Hy	Bi	Ge