



Brief article

## Songs as an aid for language acquisition

Daniele Schön<sup>a,\*</sup>, Maud Boyer<sup>b</sup>, Sylvain Moreno<sup>a</sup>,  
Mireille Besson<sup>a</sup>, Isabelle Peretz<sup>c</sup>, Régine Kolinsky<sup>b</sup>

<sup>a</sup> INCM-CNRS & Université de la Méditerranée, 31 Ch Joseph Aiguier, 13420 Marseille, France

<sup>b</sup> Unité de Recherche en Neurosciences Cognitives, ULB CP191, Avenue F.D. Roosevelt 50,  
1050 Bruxelles, Belgium

<sup>c</sup> Département de Psychologie, Université de Montréal, C.P. 6128, Montréal, Que., Canada H3C 3J7

Received 13 July 2006; revised 24 February 2007; accepted 10 March 2007

---

### Abstract

In previous research, Saffran and colleagues [Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928; Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35, 606–621.] have shown that adults and infants can use the statistical properties of syllable sequences to extract words from continuous speech. They also showed that a similar learning mechanism operates with musical stimuli [Saffran, J. R., Johnson, R. E. K., Aslin, N., & Newport, E. L. (1999). Abstract Statistical learning of tone sequences by human infants and adults. *Cognition*, 70, 27–52.]. In this work we combined linguistic and musical information and we compared language learning based on speech sequences to language learning based on sung sequences. We hypothesized that, compared to speech sequences, a consistent mapping of linguistic and musical information would enhance learning. Results confirmed the hypothesis showing a strong learning facilitation of song compared to speech. Most importantly, the present results show that learning a new language, especially in the first learning phase wherein one needs to segment new words, may largely benefit of the motivational and structuring properties of music in song.

© 2007 Elsevier B.V. All rights reserved.

---

\* Corresponding author. Tel.: +33 4 91 16 41 30; fax: +33 4 91 16 44 98.  
E-mail address: [schon@incm.cnrs-mrs.fr](mailto:schon@incm.cnrs-mrs.fr) (D. Schön).

*Keywords:* Statistical learning; Language; Music; Song

---

## 1. Introduction

Songs are ubiquitous. Yet, their function and origin remain largely unknown. The myth of Orpheus, whose songs could charm wild beasts and coax even rocks and trees into movement, illustrates the power of songs. The emotional appeal of songs and of music in general, is the most widely accepted explanation of their function. Similarly, children's songs may have an emotion regulation function, as does infant-directed speech (Trainor, Austin, & Desjardins, 2000), but they may also subserve a linguistic function (Trainor & Desjardins, 2002).

Indeed, songs may contribute to language acquisition in several ways. First, the emotional aspects of a song may increase the level of arousal and attention. Second, from a perceptual point of view, the presence of pitch contours may enhance phonological discrimination, since syllable change is often accompanied by a change in pitch. Third, the consistent mapping of musical and linguistic structure may optimize the operation of learning mechanisms.

One of the first challenges in learning a new language is to segment speech into words (Jusczyk & Aslin, 1995). Each of us will encounter such difficulty when trying to learn a foreign language which at first sounds like an uninterrupted stream of meaningless sounds. Indeed, word boundaries are not necessarily marked by consistent acoustical cues such as pauses or accents. In addition, even if there were reliable acoustical cues to word boundaries, some minimal lexical knowledge would be a prerequisite for their efficient use (Perruchet & Vinter, 1998). Yet, the distributional statistics of sub-word units provide sufficient information to discover word boundaries, at least in the very first steps of language acquisition. These distributional statistics are *transitional probabilities*, that is the probability of an event Y happening given event X. Generally, syllables that follow one another within a word will have a higher probability than syllables that are at word boundaries. For instance, given the phonological sequence *prettybaby*, the transitional probability is greater from *pre* to *ty* than from *ty* to *ba*.

Both adults and infants can use these statistical properties of syllable sequences to extract words from continuous speech (Saffran, Aslin, & Newport, 1996a, 1996b; see Yang, 2004 for a different point of view). Moreover, this statistical learning ability is not only language-related, but can also operate with non-linguistic stimuli such as tones (Saffran, Johnson, Aslin, & Newport, 1999). These findings suggest a similar statistical learning mechanism for speech and tone sequences segmentation, thereby raising the intriguing possibility that a common learning device may be involved for both language and music. This possibility is supported, from an evolutionary perspective, by an intricate evolution of the language and music faculties (Besson & Schön, 2003; Brown, 2001; Hauser, Chomsky, & Fitch, 2002), and, from a neuroscientific perspective, by recent findings showing that neural networks subserving lan-

guage and music perception are partly overlapping (Koelsch & Siebel, 2005; Maess, Koelsch, Gunter, & Friederici, 2001).

In the studies presented hereafter we compared learning based on spoken sequences to learning based on sung sequences. We hypothesized that, compared to speech sequences, a consistent mapping of linguistic and musical information (song) would enhance learning.

## 2. Experiment I

### 2.1. Participants

In the first experiment, 26 native French participants (mean age 23) listened to seven minutes of a continuous speech stream resulting from the concatenation of six three-syllables nonsense words (hereafter words) that were repeated in a pseudo-random order.

Note that this experiment is almost identical to that of Saffran and colleagues (Saffran, Newport, & Aslin, 1996b), except that participants listened to 7 min instead of 21 min of speech stream, therefore considerably reducing learning time. In previous testing with 21 min exposure periods, we were able to replicate Saffran et al.'s results, thereby showing that the actual linguistic material could be learned (performance being significantly different from chance,  $p < 0.01$ ). The choice of reducing learning time was based on the hypothesis that while a period of seven minutes might not be sufficient for learning from speech sequences, it would be sufficient for learning from sung sequences.

### 2.2. Material

The structure of the continuous stream of speech was identical to that used by Saffran et al. (1996b). Two minor changes were made. First we adapted consonants and vowels to French. Secondly we used a more recent synthesizer in order to generate the sound sequence. The way we created the stimuli was to replace each consonant and each vowel of Saffran's stimuli with one and only one consonant and vowel (ie using a simple one to one matching). The language consisted of four consonants and three vowels, which were combined into a set of 11 syllables. These were then combined to give rise to six trisyllabic words (gimysy, mimosi, pogysi, pymiso, sipygy, sysipi). Transitional probabilities within words ranged from 0.31 to 1.0. By contrast transitional probabilities across word boundaries were lower and ranged from 0.1 to 0.2. This was obtained by combining in a random order 108 repetitions of each of the six words, with the only constraint of never repeating the same word twice in a row. The text was then synthesized using the Mbrola speech synthesizer (<http://tcts.fpms.ac.be/synthesis/mbrola.html>). It is important to note that in using such a procedure, no acoustic cues were inserted at word boundaries. The result was a rather monotone and continuous stream of syllables.

### 2.3. Procedure

During the learning phase, participants were told they would listen for several minutes to a continuous stream of syllables (spoken or sung depending on the experiment). They were asked to listen carefully to these sounds, without trying to analyse them. During the testing phase, subjects were asked to indicate by pressing one of two buttons on a computer keyboard which of two strings (always presented auditorily) was most likely to be a word from the language. Each test trial consisted of two trisyllabic items: while one item was a “word” from the nonsense language, the other was not (hereafter, *part-word*). Part-words were constructed with the same syllable set as words (gysimi, mosigi, pisipy, pygyimi, sogimy, sypogy). Each word of the language was presented with each part-word, making up 36 pairs. The part-words consisted of either the last syllable of a word, plus the first syllable pair of another word, or the last syllable pair of a word plus the first syllable of another word. For instance, the last syllable of the word sysipi was combined with the first two syllables of sipygy to create the part-word pisipy. All stimuli were presented using headphones.

### 2.4. Results

The results of Experiment 1 showed that the participants’ level of performance was not significantly different from chance (48% correct,  $p = 0.45$ ; see Table 1 and Fig. 1). After 7 min of exposure, they were not able to discriminate words from part-words.

## 3. Experiment II

The second experiment was identical to the first except that the syllables of the continuous stream were sung by the synthesizer rather than spoken. Twenty-six new native French participants (mean age 23) took part in this experiment. Note that the testing phase was also identical to the previous experiment, insofar as we used spoken items and not sung items. We did this because we wanted to test only lan-

Table 1  
Percentages of correct responses and standard deviation for each group (diagonal)

	Exp 1 Speech	Exp 2 Song	Exp 3 Song	Chance
Exp 1 Speech	48% (sd. 12)			$p = 0.45$
Exp 2 Song	$p < 0.0001$	64% (sd. 10)		$p < 0.0001$
Exp 3 Song d.	$p = 0.022$	$p = 0.029$	56% (sd. 10)	$p = 0.005$

Statistical comparison (bilateral *t*-test) of each experimental group versus chance (right column), and of each group versus the other two groups ( $p$  values of post hoc tests are corrected for multiple comparisons, Tukey test).

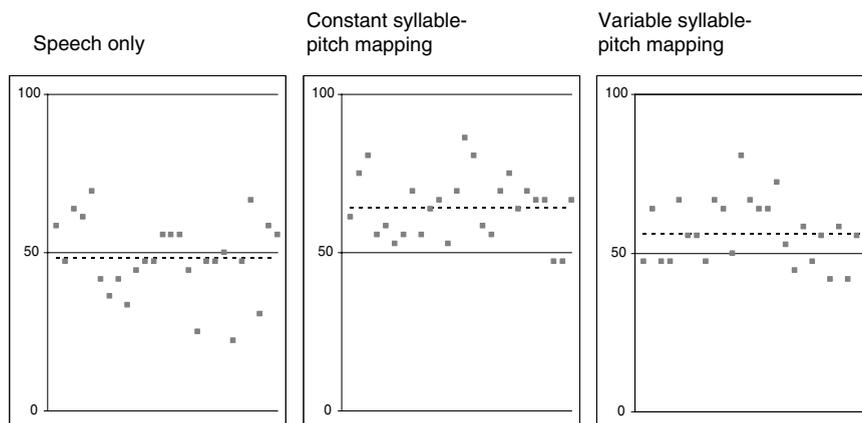


Fig. 1. Percentages of correct responses for each subject in the three different experimental groups. The dashed line indicates the average value within each group.

guage learning and we did not want to interfere with musical information in the testing phase.

### 3.1. Material

Each of the 11 syllables was associated to a distinct tone (see Fig. 2). The eleven tones were C5, D5, F5, G5, A5, B5, C6, Db6, D6, E6, and F6. Each tone was combined with a specific syllable. Therefore each word was always sung on the same melodic contour (gimysy C5 D5 F5, mimosi E6 Db6 G5, pymiso B5 E6 F6, pogysi D6 C5 G5, sipyyg G5 B5 C6, sysipi F5 G5 A5). Note that, although we intentionally chose the syllable-pitch association in such a way to have no contour change within each word, subjects could not segment on the basis of such a feature. For instance word sysipi rising F5G5A5, could be followed by a higher pitch, eg mimosi E6Db6G5 (thus still rising contour) or by a lower pitch, eg sipyyg G5B5C6 (change in contour). This continuity in contour at word boundaries took place in 12 out of 30 word transitions. Also note that the mean pitch interval within words was not significantly different from the mean interval between words, although there was a tendency of the latter of being larger (3.3 vs 5.5 half tones in average).

The sound stream was synthesized in the same way as in the previous experiment, but using precise pitch information for each syllable. Consequently, the statistical structure of syllables and tones in the sung continuous stream was identical and superimposed, with constant syllable-pitch mapping.

### 3.2. Results and discussion

Unlike in Experiment 1, participants did learn the words (64% correct,  $p < 0.0001$ ; see Table 1 and Fig. 1). Therefore, the simple addition of musical information allowed participants to discriminate words from part-words.

How may language learning benefit from musical information? As pointed out above, this may happen in at least three different ways. First, a general increase in the level of arousal or attention might increase overall performance. Second, the presence of tonal and discrete pitch changes between syllables may enhance phonological boundaries and therefore increase phonological discrimination. Indeed, syllables may be distinguished not only on the basis of their phonetic properties, but also on the basis of pitch information, and may also benefit of the gestalt properties of pitch, especially of grouping. Third, the consistent mapping of linguistic and melodic boundaries may enhance global transitional probabilities, thereby increasing the efficacy of the statistical learning mechanism.

In order to sort out which of these possible explanations best explains the effect of music facilitation, we conducted a third experiment.

#### 4. Experiment III

While the syllable sequences were still sung, linguistic and musical boundaries no longer matched (variable syllable-pitch mapping). This allowed us to (1) keep arousal constant because music had exactly the same structure as in Experiment 2, and (2) preserve phonological boundaries enhancement, because each syllable was still sung on a precise pitch. However, by decorrelating linguistic and musical boundaries, we eliminated the superposition of linguistic and melodic transitional probabilities. If we were to find the same facilitation effect as in the second experiment, then the effect should be due to arousal/attention or boundary enhancement. By contrast, if the effect were to disappear, then it would mostly be due to superposition of transitional probabilities. Again, 26 native French participants took part in the experiment (mean age 23.5).

##### 4.1. Material

The statistical linguistic and musical structure were preserved. However, they were not in phase any more. Therefore, word and pitch boundaries did not occur at the same time. More precisely, musical boundaries were shifted of one step to the right (i.e. one syllable later; see Fig. 2). Thus, while the second and third syllables of each sung word had consistent pitches, the first syllable could be sung on six different pitches (in a random manner). For instance, the word *pymiso* could be sung on F5 B6 F6 or on G5 B6 F6, depending on whether the preceding word was *gimysy* (originally C5 D5 **F5**) or *mimosi* (originally E6 Eb6 **G5**). Also note that the mean pitch interval within words was not significantly different from the mean interval between words, although there was a very weak tendency of the latter of being larger (3 vs 4.3 half tones).

Note that the testing phase was also identical to the previous experiments, insofar as we used spoken items and not sung items.



Fig. 2. Illustration of the experimental material used in Experiments 2 and 3. In the constant syllable-pitch mapping (Exp. 2) each syllable is always sung on the same pitch (e.g. “py” on B5) and the pitches used for a given word are always the same (e.g. “pymiso” – B5 E6 F6). In the variable syllable-pitch mapping (Exp. 3) one can see that these properties do not hold anymore (e.g. “sy” and “pymiso”, dashed lines).

#### 4.2. Results and general discussion

Participants’ level of performance was significantly different from chance (56% correct,  $p < 0.005$ ) and stands exactly in between that obtained in the two previous experiments (see Table 1 and Fig. 1). A one-way ANOVA including the three groups showed a highly significant main effect ( $F(2, 75) = 14.1$ ;  $p < 0.0001$ ). Post hoc tests (Tukey test) showed that the percentage of correct responses in Experiment 3 was significantly higher than in Experiment 1 with linguistic training only ( $p = 0.022$ ), and was significantly lower than in Experiment 2 with sung training with overlapping linguistic and musical structures ( $p = 0.029$ ).

This series of experiments allows us to separate the role of redundant statistical structure and perceptual saliency in learning a novel language. Indeed, the finding that the level of performance in the variable syllable-pitch mapping condition (Experiment 3) is lower than in the constant syllable-pitch mapping condition (Experiment 2) implies that superposition of transitional probabilities does play an important role in learning. At the same time, the finding that the level of performance in the variable syllable-pitch mapping condition (Experiment 3) is higher than in the speech only condition (Experiment 1) implies that arousal and/or boundary enhancement also play a role in learning. This is in line with previous results with babies showing that infant-directed speech increases infants’ attention to fluent speech and consequently to the statistical relationship between syllables (Thiessen, Hill, & Saffran, 2005). Moreover, if we were to consider that music is akin to prosody, these results would be in line with previous findings showing that prosodic information is important for segmentation. Although we did not add proper prosodic cues, such as lengthening, that have been used in previous studies (e.g. Saffran et al., 1996b), the fact of adding melodic information might facilitate grouping by means of gestalt properties and this may result in enhanced segmentation. Further studies are needed in order to see what are the effects of tonal and/or contour properties of music in facilitating segmentation.

Another interesting finding is that our results seem to point to the fact that, in the presence of multiple statistical cues, linguistic statistical cues take precedence over musical statistical cues. Indeed, if participants were to rely more on musical

statistical structure when the syllable to pitch mapping is variable (Experiment 3), they would not have succeeded in the word identification test. Such a linguistic predominance would not be surprising, insofar as our participants were all adult nonmusicians. Different results might be found with musicians, or with infants, for whom prosodic cues are not only relevant (e.g. Nazzi, Bertoncini, & Mehler, 1998; Ramus, Nespore, & Mehler, 2000; Weber, Hahne, Friedrich, & Friederici, 2004), but can even be more important than statistical cues (Johnson & Jusczyk, 2001; Kuhl, 2004). Nonetheless, this experimental design cannot truly claim whether learners were relying more on the music or the language, because only the language learning was tested. Further studies will be necessary, namely testing test learning of the tone-words rather than the speech-words.

It is important to note that the extraction of transitional probabilities extends well beyond language and music, in domains that have neither linguistic nor musical roots, including the visual domain (Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002). Thus, the results presented here, showing facilitation of musical information on language learning, may be taken to reflect the fact that redundant information in general is easier to process, not only across the linguistic and musical domains, but more generally throughout cognitive domains. Indeed, the intersensory redundancy hypothesis (Bahrick & Lickliter, 2000) holds that information presented redundantly and in temporal synchrony across two sense modalities selectively recruits attention and facilitates perceptual differentiation more effectively than does the same information presented unimodally. However, what may be specific about music and language in song is that they share the same modality and allow a rather unique overlap of spectral and temporal information. Such a combination may be more efficient than a combination across sensory modalities. More experiments are needed to investigate whether the results really pertain to a special relationship between music and language or are rather linked to intersensory redundancy in general.

Overall, our results are clear in pointing to the fact that learning is optimal when the conditions for both the emotional/arousal and linguistic functions are fulfilled. Therefore, learning a foreign language, especially in the first learning phase wherein one needs to segment new words, may largely benefit from the motivational and structuring properties of music in song. Whether this learning gain will extend to language acquisition in infant would be interesting to explore in future work. Indeed, if it were the case, it would support the idea that lullabies and children's songs may have not only an emotional (communicative and reassuring) function, but would also facilitate linguistic processing due to their simple and repetitive structure.

### **Acknowledgements**

The series of experiments reported above were conducted thanks to the support of the Human Frontier Science Program RGP 53/2002 "An interdisciplinary approach to the problem of language and music specificity". We thank Johannez Ziegler, José

Morais, Reyna Gordon and one anonymous reviewer for helpful comments on previous versions of this manuscript.

## References

- Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, *36*, 190–201.
- Besson, M., & Schön, D. (2003). Comparison between language and music. In I. Peretz & R. Zatorre (Eds.), *Neurosciences and music* (pp. 269–293). Oxford: Oxford University Press.
- Brown, S. (2001). Are music and language homologues?. *Annals of the New York Academy of Sciences* *930*, 372–374.
- Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 15822–15826.
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve?. *Science* *298*, 1569–1579.
- Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*, 1–20.
- Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*, 1–23.
- Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*, B35–B42.
- Koelsch, S., & Siebel, W. A. (2005). Towards a neural basis of music perception. *Trends in Cognitive Science*, *9*, 578–584.
- Kuhl, P. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews. Neuroscience*, *5*, 831–843.
- Maess, B., Koelsch, S., Gunter, T. C., & Friederici, A. D. (2001). Musical syntax is processed in Broca's area: An MEG study. *Nature Neuroscience*, *4*, 540–545.
- Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology. Human Perception and Performance*, *24*(3), 756–766.
- Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*, 246–263.
- Ramus, F., Nespors, M., & Mehler, J. (2000). Correlates of linguistic rhythm in the speech signal. *Cognition*, *75*(1), AD3–AD30.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.
- Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.
- Saffran, J. R., Johnson, R. E. K., Aslin, N., & Newport, E. L. (1999). Abstract Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27–52.
- Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant directed speech facilitates word segmentation. *Infancy*, *7*, 49–67.
- Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion?. *Psychological Science* *11*, 188–195.
- Trainor, L. J., & Desjardins, R. N. (2002). Abstract Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychonomic Bulletin & Review*, *9*, 335–340.
- Weber, C., Hahne, A., Friedrich, M., & Friederici, A. D. (2004). Discrimination of word stress in early infant perception: Electrophysiological evidence. *Brain Research. Cognitive Brain Research*, *18*(2), 149–161.
- Yang, C. (2004). Universal Grammar, statistics or both?. *TICS* *8*, 451–456.