Modeling the Acquisition of Statistical Regularities in Tone Sequences

Amaury Hazan (ahazan@iua.upf.edu), Piotr Holonowicz (pholonow@iua.upf.edu) ,
Inês Salselas (isalselas@iua.upf.edu), Perfecto Herrera (pherrera@iua.upf.edu),
Hendrik Purwins (hpurwins@iua.upf.edu)
Universitat Pompeu Fabra

Alicja Knast (alicja.knast@plymouth.ac.uk), Simon Durrant (simon.durrant@plymouth.ac.uk)
University of Plymouth

Abstract

Sequence learning is an important process involved in many cognitive tasks, and is probably one of the most important processes governing music processing. In this work we build and evaluate computational models addressed to solve a tone-sequence learning task in a framework which simulates forced-choice tasks experiments. The specific approach we have selected is that of Artificial Neural Networks in an on-line setting, which means the network weights are always updated when new events are presented.

Here, we aim at simulating the findings obtained by Saffran, Johnson, Aslin, and Newport (1999). We propose a validation loop that follows the experimental setup that was used with human subjects, in order to characterize the networks' accuracy to learn the statistical regularities of tone sequences. Tone-sequence encodings based on pitch class, pitch class intervals and melodic contour are considered and compared. The experimental setup is extended by introducing a pre-exposure forced-choice task, which makes it possible to detect an initial bias in the model population prior to exposure. Two distinct models (i.e. Simple Recurrent Network or a Feedforward Network with a time window of one event) lead to similar results. We obtain the most consistent learning behavior using an encoding based on Pitch Classes, which is not a relative representation. More importantly, our simulations and additional behavioral experiments highlight the impact of tone sequence encoding in both initial model bias and post-exposure discrimination accuracy. Furthermore, we suggest that melodic encoding and representation should be further investigated when inspecting and modeling behavioral experiments involving musical sequences.

Keywords: statistical learning; computational modeling; music

Introduction

In the context of the EmCAP Project [1] we are developing a computational model of musical memory and expectation that would form a module of an interactive music system. The model should take as input a musical stream and form specific expectations regarding the future sequence events, based on the sequence listened so far, and the internal representations developed by learning. As an outcome the model must be able to (1) Expect musical sequences based on what it has been exposed to, (2) Represent aspects of the structure of the attended material in a meaningful way from the cognitive point of view. While the first outcome of the system is closely related to attempts of building interactive music systems (Pachet, 2003), our second aim constrains the model to be informed of music cognition findings as observed in psychological experiments. In our view, a main concern lies in finding specific musical and cognitive tasks that a model should perform to be validated. In this perspective, we exploit the idea that statistical environmental regularities are fundamental for perception and knowledge acquisition. Statistical regularities in the environment influence the processing of information in the brain, such as learning, memory and inductive inference. So far, humans' exploitation of these statistical regularities in cognitive mechanisms has been subject of study by cognitive sciences (Barlow, 2001). Learning, this way, could be seen as the internalization of environmental regularities. Gestalt laws (proximity, common fate, good continuation), that play an important role in current psychological research, can also be seen as statistical inferences from environmental regularities. Learners take advantage of statistical information of syllable sequences such as the distribution of patterns of sounds to discover word boundaries. Several works have devised an experimental protocol for assessing how humans learn regularities in acoustic sequences, made of either tones, phonemes, or timbres. The regularities observed in the sequence can be derived from transition probabilities (Saffran, Newport, & Aslin, 1996; Saffran et al., 1999; Tillmann & McAdams, 2004), finite-state automata (Loui, Wessel, & Hudson Kam, 2006), or grammars (Bigand, Perruchet, & Boyer, 1998). While these experiments can be seen as a means to validate models of expectation, we also suggest that computational simulations may be used to inspect further and eventually validate the experimental protocols themselves. As a starting point, we aim at simulating the experiment presented by Saffran et al. (1999). This latter consists of a non-linguistic analogue of the word segmentation task applied to tone sequences. As we will see in next section, the structures to be implicitly learned depend on the transition probabilities between tones in the attended stream. Because of the very statistical nature of these tone sequences, simulating the learning process using statistical models may be seen as straightforward because the learned models would eventually reflect the statistical regularities from which the stimuli were created. However our experimental results suggest that this is not always the case and that the outcome depends on the tone sequence encoding, the statistical model, and the specific task.

[1] EmCAP (Emergent Cognition through Active Perception) is a European Commission research project. http://emcap.iua.upf.es/

Background

Transition statistics of tones

Saffran et al. (1999) focused on assessing whether humans can learn regularities related to the transition probabilities regulating the elements inside words or the word transitions. The authors created a set of artificial stimuli by setting high inside-word and low across-boundaries transition probabilities. In this work, two languages L1 and L2 were created. Each one contained 6 tone triplets, called tone-words. First, a random sequence of words of the defined language was presented to the subjects. However the tone triplets were presented in a regular order. There was no explicit cue indicating the boundaries among them. This means that the presented material appeared as a stream of tones which could be only segmented using the statistical regularities of words. In the first experiment, words from L1 were *non-words* in L2, and vice versa. That is, there was no word in one language that appeared, even partly, in the other language. In the second experiment, words from one language were *part-words* in the other language, that is, only one tone differed between each language word. After exposure, the subjects had to perform forced-choice tasks involving exhaustive word-pairs belonging to each language. The task consisted in choosing which word of the pair had been effectively heard in the presented material. This study pointed out that the subjects are able to categorize above chance the words belonging to the material they were exposed. This means that the subjects are able to segment the input stream into words and to distinguish if a word presented subsequently belongs to the sequence they have been exposed to. Also, because in Experiment 2 the words from each language were more similar, the categorization accuracy of subjects is lower than in Experiment 1. The tone words used in Experiment 1 and 2 are given in Table 1.

Table 1: Tone words used in (Saffran et al., 1999)

Experiment	Language 1	Language 2
Exp. 1	ADB,DFE,GG#A, FCF#,D#ED,CC#D	AC#E,F#G#E,GCD#, C#BA,C#FD,G#BA
Exp. 2	ADB,DFE,GG#A, FCF#,D#ED,CC#D	G#DB,DFF#,FG#A, C#CF#,D#EG#,CC#B

Artificial Neural Networks for statistical sequence learning

Here we use Artificial Neural Networks (ANN) to learn to predict the continuation of an encoded tone sequence, based on the tones observed so far. Two types or ANN are considered here, namely Feed-forward Neural Networks (FNN) and Simple Recurrent Networks (SRN). In the FNN, inputs are presented to an input layer and successively transformed and propagated into successive layers via connection weights until activating the output layer. A learning rule such as back-propagation (Rumelhart & McClelland, 1986) is applied to update the connection weights of the network according to the measured mismatch. FNN can be applied to *next-event prediction* tasks by using as inputs a certain number of past events and using as output the next event to be predicted (Kuhn & Dienes, 2008). The number of past events applied to the input layer determines the context available to provide a prediction. SRN are a variation on the FNN. A three-layer network is used, with the addition of a set of "context units" in the input layer. There are connections from the middle (hidden) layer to these context units fixed with a weight of one. The fixed back connections result in the context units always maintaining a copy of the previous values of the hidden units (since they propagate over the connections before the learning rule is applied). Thus the network can maintain a sort of state, allowing it to perform such tasks as sequence-prediction that are beyond the power of a standard feed-forward neural network. Simple Recurrent Networks have already been used in several works in order to build computational models of sequence learning from a cognitive perspective (Elman, 1990; Cleeremans & McClelland, 1991). Indeed these techniques have a great potential for modeling sequence processing, mainly because they do not rely on a specific time-window as used in FNN. There are other more recent options for modeling musical sequence processing from a statistical point of view, such as Echo State Networks (Jaeger, 2003), Self-Organizing Maps (Tillmann, Bharucha, & Bigand, 2000), N-grams (Pearce & Wiggins, 2004) or Bayesian Networks (Temperley, 2006). However these models are not addressed here and are left for further investigation.

Simulation setup

In this section, we provide details about our experimental setup, the alternatives we use in order to encode tone sequences, and our ANN models settings. First, we present in Figure 1 an overview of the experimental setting used in both original experiment and our simulation.

Tone sequence encoding

- Pitch Class (PC): each tone is encoded using a pitch class representation: we use 12 input units, for representing a given pitch we set one unit to one while the others are set to zero.

- Pitch Class Intervals (PCI): Each interval from one tone to the next one is encoded using a pitch class representation: we use 25 input units, for representing a given interval we set one unit to one while the others are set to zero. The 25 units allow to represent intervals ranging from -12 to +12 semitones.

- Melodic Contour (C): Each interval from one tone to the next one is encoded using a contour representation: we use three input units, for representing a given interval we set one unit to one while the others are set to zero. The three units allow to represent the contours *down*, *same* and *up*.

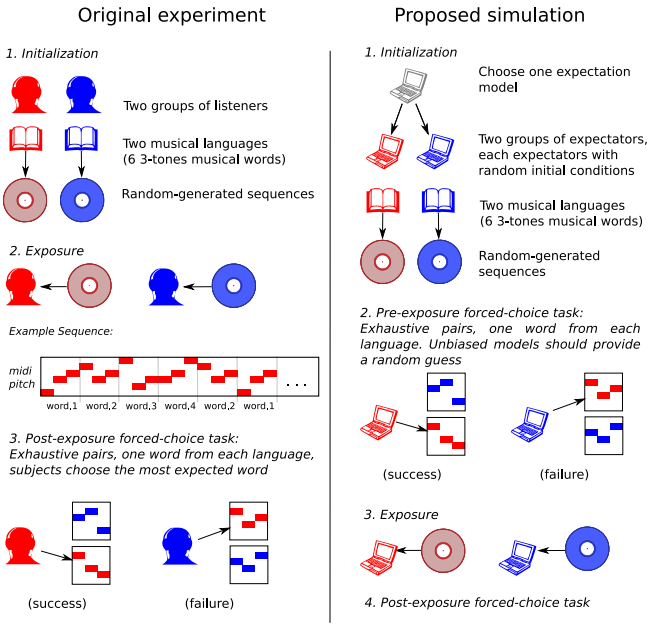


Figure 1: Overview of the experimental setup. left: original experiment, right: simulation

ANN settings

ANN are usually trained in several epochs. Then a test phase, in which no weight update takes place, is subsequently performed. Here, by analogy with the approach of Kuhn and Dienes (2008), we make no distinction between training and test mode. At each time step, even during the forced-choice task, the network weights are updated to reduce the mismatch between their expectation and the next note event. The number of epochs is set to 1, because we want to reproduce a psychological experiment in which the subjects attend the sequence of stimuli only once. We use an FNN with a time window of one event, that is, the FNN network has only access to the current event which predicting the next one. For both FNN and SRN, the detail of the parameters we explored is given below.

Exploration of the model parameters For our experiments, we used a set of parameters for defining and training the SRN. These parameters are learning rate, momentum, and number of hidden (and context) nodes. As a comparison with (Kuhn & Dienes, 2008), we do not allow a very large number of hidden and context units, for instance 60 or 120. This is because we believe that the task addressed by (Saffran et al., 1999) involves smaller time dependencies than the bi-conditional learning task addressed by Kuhn and Dienes (2008). Moreover we think that this may reduce the risk of structural overfitting. Thus, we use a smaller number of hidden unit. We summarize the set of possible parameters in Table 2.

Table 2: Parameter Set for the SRN

Parameters	Values
learning rate	0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9
hidden units	2, 5, 10, 15, 30

Simulating the forced-choice task

In order to model the forced-choice task we compare, word tone or interval (depending on the selected coding schema), the model predictions with the actual next tone or interval. The word from which the lowest mismatch is observed is selected as the chosen word. Figure 2 shows how the forced-choice task is simulated for either interval-based encodings or tone-based encodings.

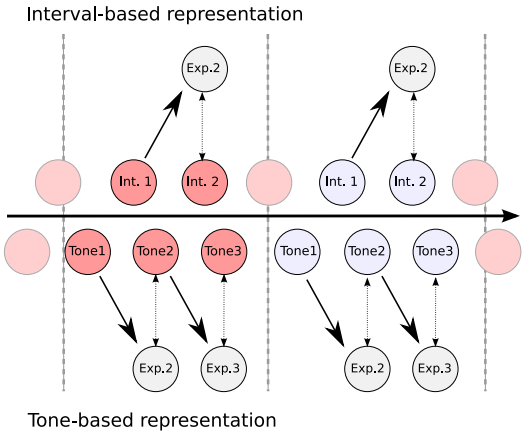


Figure 2: Forced-choice task simulation. The horizontal axis indicates time. Bottom: tone-based encoding, circles represent successive tones. Top: interval-based encoding, circles represent successive intervals. Plain diagonal arrows show models producing expectations of the next event. Only the events which are involved in the mismatch computation are labelled. Vertical bidirectional arrows show from which prediction the mismatch is measured. Dashed vertical lines show word boundaries.

Experimental loop

We run our simulations using the following general loop.

1. Create, for each language, a random sequence of tones by concatenating words from the corresponding language. Following Saffran et al. (1999), we create for each language 6 blocks of 18 words each by randomly picking words from this language. Words never appear twice in a row. Then, the blocks are concatenated randomly to form sequences equivalent to a 21 minutes auditory stream.

2. Create two network instances for simulating an individual from Group 1 and another from Group 2. Both networks have initially random weights and activations

3. For each network, perform the forced-choice categorization task on all possible combinations of L1-L2 words. Store the recognition accuracy before exposure. During this task, the order of presentation of each stimulus pair is random. Moreover, when each pair is presented, the words from the two languages are presented in random order.

4. Present to each network its corresponding sequence. This is the exposure step.

5. For each network, perform again the forced-choice categorization task. The settings are similar to those presented in step 3. Store the recognition accuracy after exposure.

We repeat this loop 100 times for each experiment in order to extract a recognition accuracy score for each instance of the network.

Results and discussion

In Figure 4 (Figure 5 respectively) we show the results of our experiments involving the SRN (respectively FNN) model. For both models, the best results were obtained using the following parameters: 2 hidden units and a learning rate of 0.01.

Influence of prediction model

The first observation to be made concerns the similarity between the results obtained in Figure 4 using the SRN model, and in 5 using the FNN model, independently of the encoding used. Our results suggest that the SRN model can not take advantage of a longer context when providing a prediction, which may confirm that the task presented in (Saffran et al., 1999) can only be solved by means of computing transition probabilities between successive events.

Pre-exposure bias

The pre-exposure scores average and standard deviation can reveal some initial bias towards one initial language, depending of the encoding used. When using the Pitch Class encoding, the pre-exposure accuracy is consistently close to the 50% baseline, which would be the outcome of a random guess for an unbiased model. However, interval-based encodings seem to exhibit a slightly higher bias, in which the pre-exposure accuracies may give preference to a particular language. This bias is statistically significant with respect to the Contour-based encoding. In this case, the pre-exposure forced-choice accuracies for Language L1 and L2 consistently belong to different distributions ($p < 0.05$), and have a bias in favor of Language 1 for Experiment 1, and Language 2 for Experiment 2. In (Saffran et al., 1999) there was no pre-exposure forced-choice task, so the information regarding a possible pre-exposure bias is not available. This is the reason why we have carried out an additional behavioral experiment (Knast, Durrant, Miranda, & Denham, 2008). This experiment reproduced the experimental setup and tone word alphabet of (Saffran et al., 1999) Experiment 1, but a pre-exposure forced-choice task was introduced. There were 24 individuals, aged 23-44 (average: 31 years old). In Figure 3, we show the distribution of selected tone words during the pre-exposure forced-choice task. The distribution presented in Figure 3 shows that words from Language L2 are more frequently choosen (54.9% in average) than words from language (45.1% in average). Furthermore, some words are affected by a strong negative (e.g. word 4) or a positive bias (e.g.words 10 and 12).

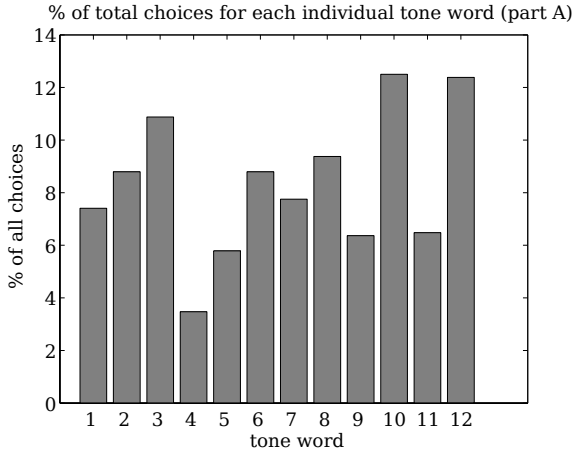


Figure 3: Pre-exposure distribution of selected tone words among Language 1 and Language 2 listeners. Tone words 1-6 belong to Language L1 and 7-12 to Language L2.

The main rationale of the original work was to investigate whether statistical learning is a domain-general mode of acquiring knowledge about environment or it is only limited to linguistics. An assumption was made that unfamiliar words build on the basis of three syllables are comparable to tone-words consisting of three notes. Here, our pre-exposure data, available either as simulations or behavioral data, suggests that there might be additional qualities of tone-word types which may have an impact on the process of learning.

Acquisition of statistical regularities

Inspecting the post-exposure results reveals distinct outcomes depending on the tone sequence encoding used. Using the Pitch Class Interval representation, the post-exposure recognition scores are higher than the baseline for Experiment 1 for both languages ($p < 0.001$). However, both SRN and FNN models fail in reproducing the results of the second experiment, because the Language L2 post-exposure score is lower than the baseline.

The Contour-based representation can not account for the results of Experiment 1, because models exposed to Language L2 exhibit a post-exposure accuracy which is lower than the baseline ($p < 0.05$). Experiment 2 is not reproduced either: in this case the model population exposed to the Language L1 exhibits a strong negative post-exposure bias towards this language ($p < 0.001$). The fact that Experiment 2 involves the comparison of words versus part-words explains well the failure in obtaining a good fit using a melodic contour

representation: indeed, the words to be discriminated during the forced-choice task are very similar when projected into a contour representation.

Overall, the most consistent improvement of the post-exposure forced-choice task accuracy for all experiments and languages ($p < 0.05$ in all cases) is obtained using a Pitch Class representation, that is, a representation where pitch is not defined with intervals. However, we were not able to reproduce the fact than Experiment 1, because it involves a comparison of words versus *non-words*, led to a higher discrimination accuracy than Experiment 2. In our simulations, the average post-exposure accuracy for Experiment 1 is 65% for Language L1 and 61% for Language L2. For Experiment 2, the average post exposure accuracy is 77% for Language L1 and 63% for Language L2.

Conclusion and future work

We have proposed in this paper an attempt towards modeling the acquisition of statistical regularities in tone sequences. We have used two Artificial Neural Network architectures to simulate the general learning trend observed by (Saffran et al., 1999). Our results show that the choice of the Artificial Neural Network architecture has little effect on the post exposure accuracy, which suggests that an extended temporal context is not necessary to model this task.

We have extended the original experiment with a pre-exposure forced-choice task and observed the outcome of this task with both simulations and a behavioral experiment. We have found that a bias towards a given language can appear, which may depend on the tone sequence representation used and enculturation effects. Therefore, we suggest that further studies aimed at investigating tone sequence learning should take into account different representations of the tone sequences and the possible initial bias listeners may exhibit.

The simulations based on interval representations such as Pitch Class Interval or Contour did not consistently account for the experimental results. However, using a tone sequence encoding based on Pitch Class, we observe, for all experiments and languages, an increase of the categorization accuracy of words versus non-words and words versus part-words in a population of prediction models after they have been exposed to tone sequences containing statistical regularities.

However, because of the specific settings and tone material used in (Saffran et al., 1999), we need to investigate further tasks to assess to which extent and in which context a given tone sequence representation is suitable. To explore further this issue, we will investigate the impact of enculturation and representation in further tone sequence learning tasks using either artificially generated material (Loui et al., 2006), and real world tone sequences (Schellenberg, 1996; Dalla Bella, Peretz, & Aronoff, 2003).

Acknowledgments

This project has been partially funded by the European Commission project EmCAP (FP6-IST, contract 013123). We thank Graham Coleman for his comments.

References

Barlow, H. (2001). The exploitation of regularities in the environment by the brain. *Behavioral and Brain Sciences*, *24*, 602-607.

Bigand, E., Perruchet, P., & Boyer, M. (1998). Implicit learning of an artificial grammar of musical timbres. *Cahiers de psychologie cognitive*, *17*(3), 577-600.

Cleeremans, A., & McClelland, J. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*(3), 235–253.

Dalla Bella, S., Peretz, I., & Aronoff, N. (2003). Time course of melody recognition: a gating paradigm study. *Perception and Psychophysics*, *65*(7).

Elman, J. (1990, apr). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Jaeger, H. (2003). Adaptive nonlinear system identification with echo state networks. *Advances in Neural Information Processing Systems*, *15*, 593-600.

Knast, A., Durrant, S., Miranda, E., & Denham, S. (2008). Enculturaton limits of the statistical learning of musical stimuli. Saffran et al. 1999 revisited. (in preparation).

Kuhn, G., & Dienes, Z. (2008). Learning of non local dependencies. *Cognition*, *106*(1), 184-206.

Loui, P., Wessel, D., & Hudson Kam, C. (2006). Adquiring new musical grammars: a statistical learning approach. In *Proceedings of the international conference on music perception and cognition.* Bolgna, Italy.

Pachet, F. (2003). The continuator: Musical interaction with style. *Journal of New Music Research*, *32*(3), 333-341.

Pearce, M. T., & Wiggins, G. A. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, *33*(4), 367-385.

Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing*. MIT Press.

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

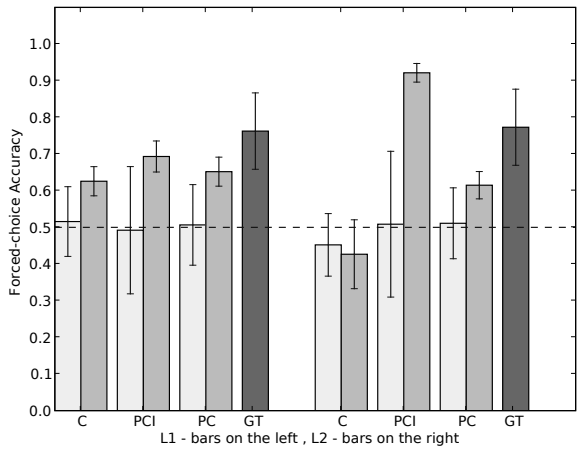
Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: the role of distributional cues. *Journal of Memory and Language*, *35*, 606 621.

Schellenberg, E. (1996). Expectancy in melody: Tests of the implication-realization model. *Cognition*, *58*, 75-125.

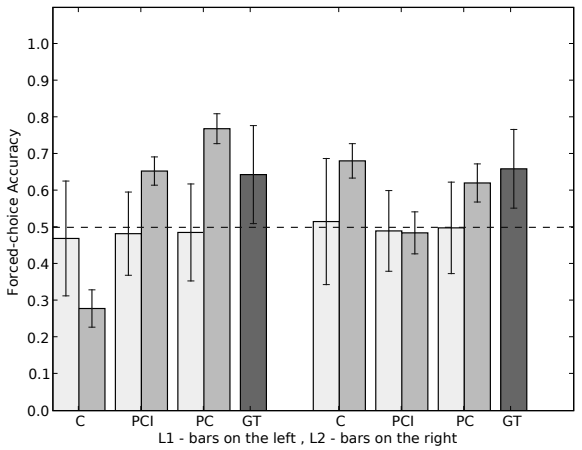
Temperley, D. (2006). A probabilistic model of melody perception. In *Proceedings of the International Symposium on Music Information Retreival (ISMIR 2006)* (pp. 276–279). Victoria, Canada.

Tillmann, B., Bharucha, J. J., & Bigand, E. (2000). Implicit learning of tonality: A self-organizing approach. *Psychological Review*, *107*(4), 885-913.

Tillmann, B., & McAdams, S. (2004). Implicit learning of musical timbre sequences: statistical regularities confronted with acoustical (dis)similarities. *Journal of experimental psychology, learning, memory and cognition*, *30*(5), 1131-1142.

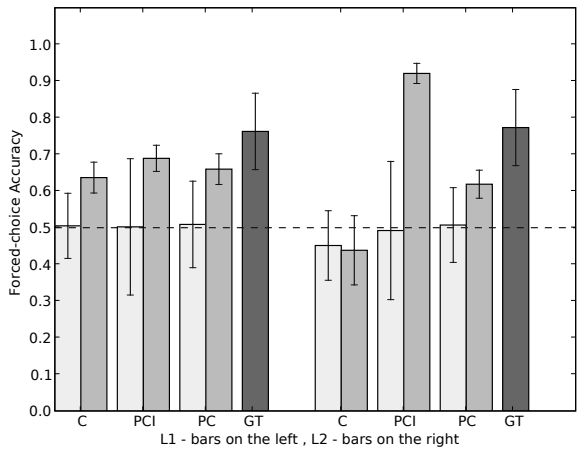


SRN model, Experiment 1

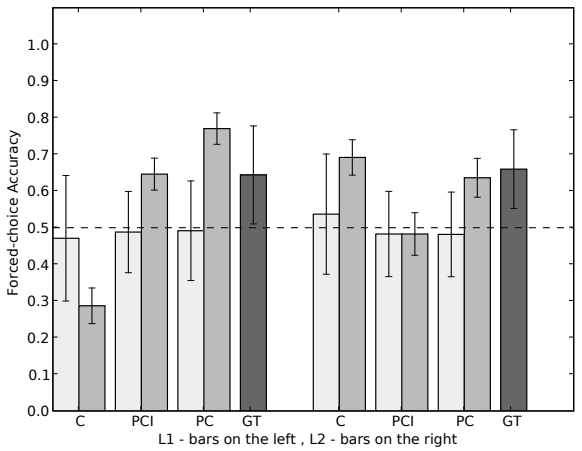


SRN model, Experiment 2

Figure 4: Forced-choice accuracy obtained with SRN predictor for distinct tone sequence encodings, compared with the subjects' response in (Saffran et al., 1999). The simulation of both Experiment 1 (left: words versus non-words) and Experiment 2 (right: words versus part-words). For each experiments and model, the results are shown for Language L1 on the left and for Language L2 on the right. For each encoding, the pre-exposure (light bars) and post-exposure (medium dark bars) mean score is plotted, along with its standard deviation over the 100 runs. Contour encoding is denoted C, Pitch Class Interval encoding is denoted PCI, and pitch class encoding is denoted PC. For each language, the right-most bar shows the ground truth post-exposure accuracy obtained by Saffran et al. (1999), denoted GT. The horizontal dashed line indicates the 50% baseline.



FNN model, Experiment 1



FNN model, Experiment 2

Figure 5: Forced-choice accuracy obtained with FNN predictor for distinct tone sequence encodings, compared with the ground truth. For the legend we refer to Figure 4.