# Acoustic analysis supports the existence of a single distributional learning mechanism in structural rule learning from an artificial language

**Okko Räsänen (okko.rasanen@aalto.fi)**
Department of Signal Processing and Acoustics, Aalto University,
Otakaari 5 A, FI-00076 Aalto FINLAND

**Heikki Rasilo (heikki.rasilo@aalto.fi)**
Department of Signal Processing and Acoustics, Aalto University,
Otakaari 5 A, FI-00076 Aalto FINLAND

## Abstract

Research on artificial language acquisition has shown that insertion of short subliminal gaps to a continuous stream of speech has a notable effect on how human listeners interpret speech tokens constructed from syllabic constituents of the language. It has been argued that the observed results cannot be explained by a single statistical learning mechanism. On the other hand, computational simulations have shown that as long as the gaps are treated as structurally significant units of the language, a single distributional learning model can explain the behavioral results. However, the reason why the subliminal gaps interfere with processing of language at a linguistic level is currently unknown. In the current work, we concentrate on analyzing distributional properties of purely acoustic representations of speech, showing that a system performing unsupervised learning of transition probabilities between short-term acoustic events can replicate the main behavioral findings without a priori linguistic knowledge.

**Keywords:** language acquisition; pattern discovery; distributional learning; acoustic analysis; lexical learning

## Introduction

There is an ongoing debate regarding the degree that distributional learning mechanisms can explain aspects of language acquisition from speech, and the degree that rule-based mental processes are required in the task (e.g., Endress & Bonatti, 2007; Laakso & Calvo, 2011; Peña et al. 2002). Experimental studies with human test subjects have shown that both infants and adults are able to learn statistical regularities in continuously spoken artificial languages and use these regularities to segment speech into word-like units (e.g., Peña et al. 2002; Saffran, Aslin & Newport, 1996). Based on these findings, it has been suggested that the listeners may be using transitional probabilities (TPs) between speech units such as phones or syllables in order to discover statistically regular segments of speech (e.g., Saffran et al., 1996). Computational simulations have also verified that the TPs between signal events can be used to discover word-like units from continuous speech, and that these units do not necessarily need to be linguistic or phonetic in nature (Räsänen, 2011).

Of especial interest is the degree that distributional learning can explain the learning of non-adjacent dependencies in a language. In earlier work, the learning of non-adjacent dependencies has been studied using an artificial nonsense language consisting of three-syllabic CVCVCV words with the middle syllable being always randomly selected from a pool of "fillers", but the first and last syllable occurring always together (hence a "*high-probability word*"). It has been found out that when human listeners are familiarized with a continuous stream of such language without gaps between the high-probability words, and then later tested for preference between three-syllabic words that have different TPs between the syllables in terms of the familiarization stream, the listeners seem to prefer words that have occurred with higher internal TPs in the familiarization stream (Endress & Bonatti, 2007; Peña et al. 2002). However, introduction of 25 ms subliminal segments of silence between the high-probability words in the familiarization stream leads to a notable change in the learning outcome: the listeners start to prefer word forms that do not necessarily have the highest TPs across all syllables in the word. Instead, the preferred words may contain partially novel surface form but have dependencies between syllables that can be explained by abstract rules that are also valid for the words in the familiarization stream (Endress & Bonatti, 2007; Peña et al. 2002).

The above finding is somewhat unexpected from the perspective of distributional learning at a linguistic level. The learning results between continuous and gapped familiarization streams should not differ as long as the perceived linguistic units and their ordering in the two conditions do not differ either. The result is also counterintuitive due to the fact that the gaps are tiny in duration in comparison to the other relevant signal segments such as syllables, and since CV-syllable based languages already contain natural silences associated with closures of plosives (e.g., word "*#pura#ki*", where # denotes a closure).

Peña et al. (2002) and Endress and Bonatti (2007) suggest that the additional silent gaps provide direct (but unconscious) cues to the segmentation of words from speech, freeing computational resources to structural learning of rule-like relations between constituents of the words. On the contrary, the absence of the gaps necessitates that the segmentation has to be first learned from the data (Endress & Bonatti, 2007; but see also discussion in Laakso & Calvo, 2011). It is therefore argued that the change in learning outcomes after introduction of the gaps provides evidence for non-distributional learning of structural relations between syllabic units (Bonatti & Endress, 2007).

However, a possible auditory processing mechanism for differentiating gaps associated with segmental cues and, e.g., the intra-word gaps related to closures of plosives has not been described in the existing work.

Lately, Laakso and Calvo (2011) have shown that the experimental results of Peña et al. (2002) and Endress and Bonatti (2007) *can* actually be modeled with a single distributional connectionist model when the silent gaps are represented as equally significant units as the consciously perceived syllables. As long as Occam's razor is concerned, the distributional model of Laakso and Calvo (2011) provides a more coherent and simple explanation for the observed data instead of resorting to the more than one mechanisms (MOM) hypothesis of Peña et al. (2002) and Endress and Bonatti (2007). However, the model of Laakso and Calvo also has a shortcoming: it does not explain how the short subliminal gaps end up with an equally large role as the syllabic units in the distributional learning process.

The goal of the current work is to study the distributional learning hypothesis in the context of the artificial language of Peña et al. (2002) by focusing on the analysis of recurring acoustic patterns in a speech stream. Unlike earlier work, we study TPs of short-term acoustic events instead of linguistically or phonetically motivated units such as syllables or segments. This provides a novel perspective to the learning problem by assuming that the listeners may not be directly analyzing the speech stream as a sequence of linguistic units, but may treat the language-learning task as a generic auditory patterning problem. Still, the current approach does not exclude the possibility that the listeners can extract basic recurring units such as syllables or segments from the acoustic speech stream and perceive these as linguistically significant units. We simply show that the behavioral results of Peña et al. (2002) and Endress and Bonatti (2007) can be explained with a single distributional learning mechanism that performs pattern discovery at the level of acoustic signal instead of assuming TP analysis of segments or syllables.

### Motivation for Acoustic Learning

There are multiple reasons to assume that the listeners may utilize generic acoustic patterning instead of purely linguistic coding of input during perception of an artificial language. First of all, test subject preferences towards specific test probe types are typically only slightly above chance level even for extended familiarization periods (Peña et al., 2002; Endress & Bonatti, 2007). If the learning would be based on categorically perceived segments or syllables, one could expect more robust preference for one probe type over another due to the systematically different overall TPs or learned rules for the tokens. Also, the initial preference for specific probe types degrades over longer familiarization periods, suggesting that the low-level distributional properties of the speech stream interfere with the processing of the abstract generalizations. Finally, the introduction of subliminal gaps introduces notable qualitative changes to the learning outcomes. Since these gaps are clearly not serving any explicit linguistic function but still affect the learning results, it can be taken as evidence that the acoustic level perception, including temporal relationships of acoustic patterns, may play an important role in the process.

Why distributional analysis at the acoustic level would then lead to different results than analysis on the segmental or syllabic level? The major difference comes from temporal relationships between sound events. At the syllabic level, the relevant units and their distances from each other are well defined. Therefore the TP statistics also become well defined after a small number of word occurrences in different contexts. At the acoustic level, a syllable is not perceived as a categorical unit with a well-defined duration, but as a constantly evolving spectrotemporal trajectory that has very low predictability over larger temporal distances. This means that the typical acoustic level dependencies are limited to a time scale much shorter than the tri-syllabic words in the artificial language of Peña et al. (2002). Therefore the acoustic TP analysis must also pay attention to dependencies at a very fine temporal resolution, potentially increasing the relative role of temporal asynchronies caused by the introduction of silent gaps to the familiarization stream.

## Material

The speech material for the experiments was reproduced from the work of Peña et al. (2002). In this material, the familiarization stream of the artificial language consists of three CV-syllable words of form $A_iXC_i$ so that each word starts with one of three possible syllables $A_i$ ($i \in \{1,2,3\}$). Importantly, the first syllable always uniquely determines the last syllable $C_i$ of the word (i.e., $P(C_i|A_i) = 1, \forall i$) so that there are also three different possibilities for end syllables. Finally, the medial syllable, or *filler*, is chosen randomly from a set of three CV syllables. In total this produces three word templates "pu … ki", "be … ga", and "ta … du" where one of the following three fillers are used in the medial position: "li", "ra" or "fo".

Based on Endress and Bonatti (2007), four types of probes were used during testing: 1) *words*, i.e., tri-syllable constructs that correspond directly to the ones used in the familiarization (e.g., $A_iXC_i$), 2) *part-words*, where the sequential order of syllables was from the familiarization data but the word straddles a word boundary (e.g., $XC_iA_j$), therefore having a smaller overall TPs across the word, 3) *rule words* of form $A_iX'C_i$, where the X' is familiar from the training but has never occurred in the word-medial position, and 4) *class words* of form $A_iXC_j$ ($i \neq j$) so that all $A_i$, X, and $C_j$ are familiar from the familiarization data but the $A_i$ and $C_j$ have never occurred in the same word (see Endress & Bonatti, 2007, for detailed word lists).

The familiarization data and test probes were synthesized into speech signals using a Kelly-Lochbaum model based articulatory synthesizer of Rasilo, Räsänen and Laine (in preparation) using articulatory positions of Finnish vowels as targets for the vowel sounds. Sampling rate of the signals was set to 16000 Hz and fundamental frequency of the

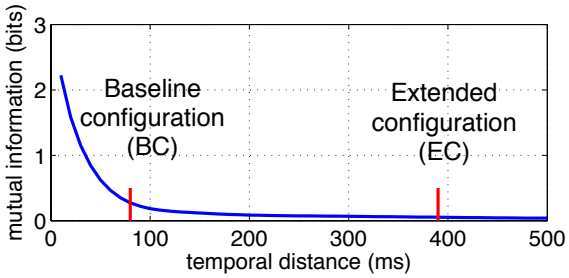


Figure 1: Temporal dependencies of acoustic events measured from continuous English speech. The two learning parameter configurations BC and EC are also shown.

speaker was set to 120 Hz. In order to create familiarization data, all words in a training epoch (one occurrence of each word) were concatenated into one long string before synthesis so that the coarticulatory effects were consistent for both intra-word and across-word transitions. In addition to the continuous stream, the gapped familiarization stream of Peña et al. was also created by inserting silent segments of 25 ms between the words. It was also confirmed perceptually that the perception of the gaps was subliminal and no other audible artifacts were introduced to the signals.

## Methods

### Preprocessing

The goal of the preprocessing was to convert synthesized speech signals into sequences of automatically discovered discrete acoustic events for further statistical modeling. This was achieved by extracting Mel-frequency cepstral features (MFCCs) from the signals using a window length of 25 ms and a step size of 10 ms (see, Appendix B in Räsänen 2011 a for detailed description). A total of 12 coefficients + energy were used. A random subset of 10000 MFCC vectors from the familiarization data set was then clustered into 64 clusters using the standard k-means algorithm. The obtained cluster centroids were treated as prototypes for the corresponding clusters ("atomic acoustic events") and each cluster was assigned with a unique integer label $i \in [1, 2, \ldots, 64]$. Finally, all MFCCs vectors were vector quantized (VQ) by representing the original feature frames with labels corresponding to the nearest cluster centroids for the given frame. This led to a signal representation where the synthesized speech was represented as a sequence of discrete elements, each element being one of the 64 possible choices and one element occurring every 10 ms.

### Discovery of Acoustic Patterns

In order to learn distributional patterns from the artificial speech data, a statistical learning mechanism is needed. In the current work, we utilized the unsupervised word learning model of Räsänen (2011) that has been shown to be able to discover recurring word patterns from real continuous speech. This algorithm will be referred to as the *unsupervised distributional learning algorithm* (UDLA).

The basic principle of the UDLA is to study the TPs between the atomic acoustic events (VQ indices) in order to discover multiple segments of speech that share similar local TP distributions. Unlike typical distributional analysis of syllabic, phonemic, or ortographic units (e.g., Saffran, 1996), UDLA analyzes TPs between short-term acoustic events at several temporal distances (lags) in parallel so that dependencies between non-adjacent acoustic events also become modeled. When recognizing novel patterns, statistical support from all lags is combined in order to provide a uniform and noise robust estimate of familiarity of the pattern. Instead of modeling global TPs, UDLA creates a separate TP model for each novel pattern discovered from the data, where a novel pattern is defined as a sequence of acoustic events whose TPs do not match any of the previously learned patterns.

From the perspective of pattern discovery, it is beneficial to study temporal dependencies up to approximately 200 ms in case of continuous speech. This is because the statistical dependencies between acoustic events diminish to a non-existent level at larger temporal distances and provide no further support for pattern discovery (Räsänen & Laine, 2012). This temporal scale also corresponds to the typical signal integration times measured in human auditory perception in the context of loudness perception or forward masking of speech sounds, suggesting that the integration times in human hearing are matched to the typical temporal structure of acoustic speech signals. As an example, Figure 1 shows the statistical dependencies of short-term acoustic events as a function of temporal distance for continuous English speech measured in terms of mutual information function (MIF; Li, 1990). As can be observed from the figure, majority of the dependencies at the acoustic level are limited to temporal distances shorter than 100 ms.

Since the amount of statistical information diminishes at longer distances, one can hypothesize that the human hearing system would be adapted to process temporal dependencies at such timescale where, on average, dependencies do exist. Therefore, in *baseline configuration* (BC), we use UDLA in a mode in which dependencies are modeled up to 80 ms, capturing approximately 90 % of the statistical dependencies in terms of MIF (Fig. 1). However, we also measure UDLA behavior in the artificial language learning task using TP modeling up to 390 ms. This configuration will be referred to as *extended configuration* (EC). In terms of the current experiments, this means that the TPs were studied at lags $\mathbf{k} = \{1, 2, \ldots, 8\}$ for BC and at lags $\mathbf{k} = \{1, 3, 5, \ldots, 39\}$ for EC, corresponding to the modeling of acoustic dependencies at temporal distances of 10 ms – 80 ms and 10 ms – 390 ms, respectively.

The hypothesis was that, if acoustic and non-linguistic patterning can explain the results of the experiment of Peña et al. (2002), and if human hearing is actually specialized for learning dependencies according the curve shown in Fig. 1, the learning outcomes in the baseline configuration should have better correspondence to the behavioral results than the extended condition. On the other hand, the extended

configuration should show higher preference for part words than class or rule words due to the diminishing role of the gaps in terms of dependencies across all temporal distances.

**Training Phase** The learning process in UDLA proceeds as follows (see also Räsänen, 2011): the sequential discrete familiarization stream $X$ is analyzed in windows of length $L_r$ elements and window step size $L_s$. For each window position, the TPs between all elements $a_i$ and $a_j$ in the window are modeled in parallel for lags $\mathbf{k} = \{k_1,k_2,\ldots k_K\}$. For the TPs in the first window position, the first statistical model $c_1$ is created by storing all transitions at all lags to a transition probability matrix. In the model, the probability of a transition from element $a_i$ to $a_j$ at lag $k$ is defined as

$$P_c^{\mathrm{S}}(a_j \mid a_i,k) = F_c(a_i,a_j \mid k) / \sum_{j=1}^{N_{\mathrm{A}}} F_c(a_i,a_j \mid k) \qquad (1)$$

where $F_c(a_i,a_j|k)$ is the frequency of ordered pairs $[a_{\mathrm{i}}\ a_{\mathrm{j}}]$ at distance $k$ in the context of model $c$.

When the window is moved incrementally across the input sequence, all previously learned models are used to recognize the contents of the current window position. First, activation $A_c(t)$ of each model $c$ at each moment of time $t$ is computed by calculating the mean of the TPs over all $\mathbf{k}$:

$$A_c(t) = \frac{1}{K} \sum_{k=1}^{K} P_c^{\mathrm{S}}(X[t] \mid X[t-k],k) \qquad (2)$$

The cumulative activation of each model is then calculated over the window and normalized by the window length:

$$A_c^{cum}(T) = \frac{1}{L_r} \sum_{x=T}^{T+L_r-1} A_c(t[x]) \qquad (3)$$

where $T$ denotes the window position. Now if activation $A_c^{cum}$ of the most activated model $c_{\mathrm{M}}$ exceeds a pre-defined familiarity threshold $t_{\mathrm{r}}$, the transition frequencies in the current window of analysis $X_T$ are used to update the statistics of the model $c_M$ according to Eq. (1). Otherwise, a new model $c_{\mathrm{N}}$ is created from the window contents using the Eq. (1). This process is repeated for the entire training data set, producing a set of models that incrementally increase their selectivity towards specific structures in the speech signal.

After the familiarization is complete, the learned models are normalized according to

$$P_c(a_j \mid a_i,k) = P_c^{\mathrm{S}}(a_j \mid a_i,k) / \sum_{m=1}^{N_C} P_m^{\mathrm{S}}(a_j \mid a_i,k) - \frac{1}{N_{\mathrm{C}}} \qquad (4)$$

where $N_{\mathrm{C}}$ is the total number of models learned. This changes the nature of the statistics so that now $P_{\mathrm{c}}$ describes how likely the given transition from $a_j$ to $a_i$ occurs in case of pattern $c$ instead of any other pattern (i.e., *classification* task). The $1/N_{\mathrm{c}}$ term forces the total activation across all models to zero at all times, ensuring that the total activation level of the system does not increase with increasing number of learned models. Note that the learning process is purely incremental and requires the storage of the previous inputs only up to maximum lag $K$ (i.e., 80 or 390 ms).

**Recognition Phase** During the testing phase, the test probes were pre-processed into discrete VQ sequences similarly to the familiarization data. Then the instantaneous activation of each model $c$ at time $t$ given input probe $X$ was measured according to

$$A_c(t) = \frac{1}{K} \sum_{k=1}^{K} P_c(X[t] \mid X[t-k],k) \qquad (5)$$

The total activation induced by the probe was then computed as

$$A_{tot} = \arg_{t,c} \max(A_c(t) \mid \forall t,c) \qquad (6)$$

In other words, the total activation caused by the probe $X$ was obtained as the maximum instantaneous activation[1] in the pool of all pattern models $c$.

## Experiments

In the experiments, UDLA was first used to discover recurring acoustic patterns from the familiarization stream, and then to recognize novel test probes using the learned models. During each test round, the system was shown one token from each of the four possible probe classes and the overall activation caused by each token was measured. A total of 600 probe quartets were generated by randomly sampling one token from each probe class for each quartet.

In all experiments, the UDLA model was run with a familiarity threshold of $t_{\mathrm{r}} = 0.16$ and window step size $L_{\mathrm{s}}$ = 50 ms (5 frames). The analysis window length was set to $L_{\mathrm{r}}$ = 200 ms and $L_{\mathrm{r}}$ = 600 ms for baseline and extended conditions, respectively, so that multiple transitions at maximal lags would fit to the analysis window. These parameters led to the learning of $N_{\mathrm{C}}$ = 26-33 acoustic patterns depending on the familiarization type (continuous vs. segmented), modeling conditions (baseline vs. extended), and on the duration of the familiarization. Since the number of learned patterns exceeded the number of unique syllables (nine), the system had learned multiple context-sensitive variants of syllable-like units.

Figure 1 shows the mean activation levels of the four different probe types (words, part words, rule words and class words) as a function of familiarization duration for segmented (top) and continuous (bottom) familiarization stream in the baseline condition with temporal dependency modeling up to 80 ms. As can be observed, the insertion of 25 ms gaps between tri-syllable words in the familiarization stream is sufficient to induce a change of preference from part words to rule words and class words. This is in line with the behavioral results of Peña et al. (2002) and Endress and Bonatti (2007) who found out that the use of subliminal

[1] The decoding criterion of probabilities was compared across numerous different possibilities, including, e.g., total activation of all models across the entire probe, temporally integrated maximum activation, and number of models exceeding a pre-defined threshold in activation. However, unlike the used approach in Eq. (6), none of the other criteria were able to replicate the main findings of Peña et al. (2002) and Endress & Bonatti (2007).

gaps in the familiarization stream causes a change of preference from part words to rule words and class words at short familiarization periods.

However, when the TPs between acoustic events are measured beyond the typical dependencies in speech signals, the situation changes notably. Figure 3 shows the mean activation levels of the probes in the extended condition where temporal dependencies are modeled up to 390 ms. Despite the fact that the only difference to the earlier simulation is the distance up to which TPs are measured, there is no sign of difference between the continuous and segmented familiarization streams.

Based on the mean probe activities, it seems that the distributional learning of acoustic patterns without any a priori or intervening linguistic component can explain the experimental results of Peña et al. (2002) and Endress and Bonatti (2007), but only if it is assumed that the system is able to learn acoustic dependencies up to a limited temporal distance defined by typical structure in continuous speech. If the dependency modeling is extended up to much longer delays, the UDLA model is no longer able to replicate the behavioral findings.

In addition to computing overall activations, pair-wise comparisons of probe activities were carried out for all possible probe pairs in the test set in order to simulate behavior in a forced-choice task similar to the one used with human experiments.

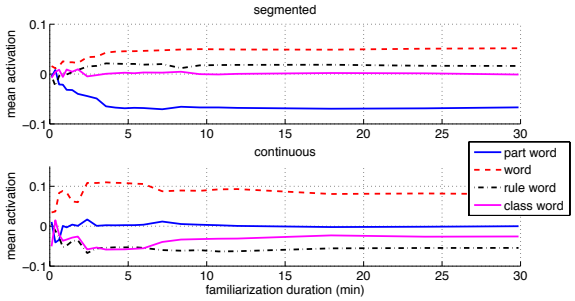


Figure 2: The mean activation levels of the four different probe types in *baseline condition* for segmented stream (top) and for continuous stream (bottom). Only relative mean activations of the probes are shown (zero mean).

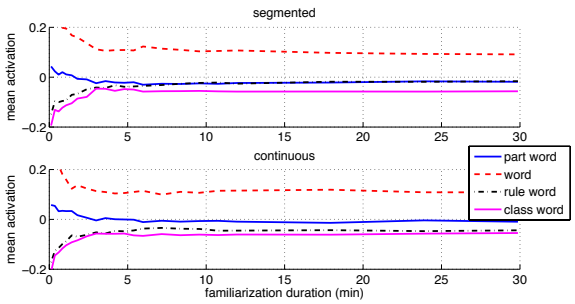


Figure 3: The mean activation levels of the four different probe types in *extended condition* for segmented stream (top) and for continuous stream (bottom).

More specifically, the relative probabilities of the tokens in each pair were compared separately across all 600 test cases in the baseline configuration. For each pair, a binary flag was used to denote a response for the probe that had the higher activation. Then the distribution of responses was tested against the null hypothesis that the model shows no preference for either probe type (*t*-test). Table 1 illustrates the results from the statistical analysis.

It is evident that the segmented familiarization stream leads to a preference order of *words* > *rule words* > *class words* > *part words* at short familiarization durations. On the other hand, continuous stream leads to order of *words* > *part words* > *rule words* and *class words*. This is largely in line with the results of Laakso and Calvo (2011), confirming that a single distributional learning mechanism can explain the change of preference between the two conditions. However, the previous studies do not always report statistically significant order of preference between all probe types (Laakso & Calvo, 2011), whereas the current simulations show statistically significant order of preference for all learning conditions except for the continuous familiarization stream of 3 minutes. This can be largely explained by the fact that the deterministic nature of UDLA leads to a consistent response pattern across multiple trials even for minor statistical biases between the probe types. In contrast, responses of human test subjects contain additional sources of variation (e.g., fatigue) and are based on a limited number of test trials, possibly rendering minor differences in probe familiarity invisible to statistical analysis.

## Discussion

In Peña et al. (2002) and Endress and Bonatti (2007) it was found that adult test subjects, when familiarized with 10 minutes of continuous stream of speech from an artificial language, prefer words over part words and show no preference between class words, part words and rule words. However, when subliminal gaps were introduced between words in the familiarization stream, the participants started to prefer class words and rule words over part words. Based on these findings, Peña et al. (2002) put forward the MOM hypothesis that the learning of a language might consist of several different processes: a distributional process responsible for discovery of statistically significant patterns and a separate mechanism responsible for modeling of structural relation between the discovered patterns. Endress and Bonatti (2007) provided further support to the MOM hypothesis by failing to replicate the behavioral findings of Peña et al. when modeling the learning task with a distributional system (a recurrent neural network or RNN).

Lately, Laakso and Calvo (2011) showed that RNNs can replicate the main behavioral findings of Peña et al. when the modeling parameters are properly set up, and when the silent gaps between syllables are modeled as separate units with equal importance to syllabic units. Their results undermine the argument for the necessity of multiple mechanisms of learning in this specific context. However, Laakso and Calvo limited their analysis to purely linguistic

Table 1: Pair-wise preference for the four different types of test probes with *segmented* (left) and *continuous* (right) familiarization streams. W stands for word, PW for part word, C for class word and R for rule word.

| | segmented | | | continuous | | |
|---|---|---|---|---|---|---|
| | preference | % | *p* | preference | % | *p* |
| 3 min | W over PW | 82.1 | 0.0000 | W over PW | 77.8 | 0.0000 |
| | R over PW | 74.0 | 0.0000 | PW over R | 57.1 | 0.0005 |
| | C over PW | 70.5 | 0.0000 | PW over C | 64.0 | 0.0000 |
| | W over R | 58.8 | 0.0000 | W over R | 89.5 | 0.0000 |
| | W over C | 68.3 | 0.0000 | W over C | 90.0 | 0.0000 |
| | R over C | 56.9 | 0.0032 | No pref. R and C | 51.4 | 0.5156 |
| 10 min | W over PW | 78.4 | 0.0000 | W over PW | 71.2 | 0.0000 |
| | R over PW | 70.0 | 0.0000 | PW over R | 60.1 | 0.0000 |
| | C over PW | 69.1 | 0.0000 | PW over C | 57.0 | 0.0006 |
| | W over R | 68.4 | 0.0000 | W over R | 83.4 | 0.0000 |
| | W over C | 74.9 | 0.0000 | W over C | 79.0 | 0.0000 |
| | No pref. R and C | 55.9 | 0.0113 | C over R | 59.7 | 0.0000 |

level, assuming that the learner perceives artificial language as a sequence of syllabic units and silences even though the silences were not consciously perceived by the participants.

Current work studied the hypothesis that the findings of Pena et al. could be based on generic distributional learning at the acoustic level instead of using linguistic level representations. More specifically, we analyzed TPs of short-term acoustic events that were extracted from speech in purely unsupervised manner. Notably, we were able to replicate the behavioral findings related to the change of preference across familiarization conditions by using the UDLA model of word learning from continuous speech, but only when the TP analysis of acoustic events was limited to a temporal window matching to the temporal dependencies of normal continuous speech (Räsänen & Laine, 2012).

If this constraint is violated by exceeding the temporal scale of modeling to several hundreds of milliseconds, the model systematically prefers words over part words, and part words over class words or rule words also in case of segmented familiarization stream. The change of model behavior is driven by the fact that the synthesized speech lacks the acoustic variability and lexical complexity of normal speech, and therefore unnaturally strong long-distance dependencies exist in the speech tokens. By modeling the TPs at increasingly long distances, the relative statistical contribution of the short-term gaps between the words in the segmented condition become too small to affect the preference of word tokens in the testing phase.

This suggests that if human responses in the task are based on acoustic level patterning, it may be the case that the human auditory system is not able to capture dependencies at extended temporal distances. This is closely related to the study of Newport and Aslin (2004) who found that adult listeners are unable to learn dependencies between non-adjacent syllables whereas dependencies between non-adjacent segments (either vowels or consonants) were readily learned when familiarized with continuous stream of artificial language. The inability to learn non-adjacent syllabic dependencies could be also explained by the finite length temporal integration in the auditory processing. Segmental dependencies with an interleaved random segment in between could be readily captured by a system modeling statistical dependencies up to, e.g., 150 ms, but dependencies across multiple syllables may simply be too distant to be captured by such short-term analysis.

Note that the inability to capture acoustic dependencies at longer temporal distances does not imply that long-range linguistic dependencies would not exist or could not be captured by a distributional learning mechanism. It is well known that such dependencies do exist. However, the huge variability and dimensionality of the acoustic space strongly points towards the necessity of an intermediate representation upon which further analysis and learning can take place. Given the current knowledge of human speech perception, it is early to say whether these units are phones, syllables, morphemes or something else (see Räsänen, 2011), and whether the computations are distributional or structural in nature. The current study does not exclude the possibility that the human listeners are directly utilizing syllable level TPs in the artificial language learning task, but simply shows that the TP analysis at the acoustic level can also explain behavioral observations to a large degree.

## Acknowledgements

This research was financially supported by Nokia NRC.

## References

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, 105(2), 247-299.

Laakso, A., & Calvo, P. (2011). How Many Mechanisms Are Needed to Analyze Speech? A Connectionist Simulation of Structural Rule Learning in Artificial Language Acquisition. *Cognitive Science*, 35, 1243-1281.

Li, W. (1990). Mutual Information Functions versus Correlation Functions. *J. Statistical Physics*, 60, 823-837.

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

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, 298(5593), 604-607.

Rasilo, H., Räsänen, O., & Laine, U. (In preparation). An approach to language acquisition of a virtual child: learning based on feedback and imitation by caregiver.

Räsänen, O. (2011). A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events. *Cognition*, 120, 149–176.

Räsänen, O., & Laine, U. (2012). A method for noise-robust context-aware pattern discovery and recognition from categorical sequences. *Pattern Recognition*, 45, 606-616.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, 274, 1926-1928.