Effects of Memory Size on Melody Recognition in a Simulation of Cohort Theory

Naresh N. Vempala*\textsuperscript{a,1}, Anthony S. Maida\textsuperscript{b}

\textsuperscript{a}Institute of Cognitive Science
University of Louisiana at Lafayette
Lafayette, LA 70504
USA

\textsuperscript{b}Center for Advanced Computer Studies
University of Louisiana at Lafayette
Lafayette, LA 70504
USA

Abstract

This research models human performance in the Dalla Bella, Peretz, and Aronoff (2003) melody recognition study. They compared performance between musicians and nonmusicians in the perception and recognition of isolated melodies. Inspired by cohort theory, they used a gating task to identify three cognitive events in the melody perception/recognition process. These events were the familiarity emergence point (FEP), the isolation point (IP), and the recognition point (RP). Dalla Bella et al. explain their results using cohort theory.

In the present research, we develop a simulation based on a sequence-recognition neural network (SRN) that is compatible with cohort theory. We use it to model the hypothesized cognitive processes underlying the occurrence of these events. Two separate networks with different stored memory sizes are used to model musicians and nonmusicians. Our simulations capture the qualitative pattern of results in Dalla Bella et al. and show how stored memory size may affect the melody recognition process. The IP is modeled using the core SRN that contains a winner-take-all (WTA) mechanism. The FEP and RP are modeled by introducing meta-level readout units that monitor the dynamically evolving state of the WTA network. The unit that models the FEP plays a causal role in determining the RP. Therefore, the FEP is not an epiphenomenon in our simulation.

It was not previously known whether the core SRN contained enough information to allow simple meta-level readout of dynamic cognitive states. Our simulations show that such readout is in fact possible.

Key words: melody recognition, cohort theory, recurrent neural network, sequence recognition.

\*Corresponding author

\textit{Email addresses:} nnv1@hotmail.com (Naresh N. Vempala), maida@cacs.louisiana.edu (Anthony S. Maida)

\textsuperscript{1}Current address: Department of Psychology, Ryerson University, 350 Victoria Street, Toronto, Ontario M5B 2K3, Canada

Preprint submitted to Elsevier

July 5, 2010
1. Introduction

Upon hearing a known melody, a listener will first experience a subjective sense of familiarity, then provisionally recognize the melody, and finally recognize the melody with a high level of confidence. This phenomenon was empirically studied by Dalla Bella, Peretz, and Aronoff (2003). Dalla Bella et al. named these subjective states the familiarity emergence point (FEP), the isolation point (IP), and the recognition point (RP), respectively. Their experiments showed that the temporal pattern of these states was somewhat different for musicians and nonmusicians. The present paper reports on a neural network modeling study of these phenomena as they occur in isolated recognition of a known melody.\(^2\)

Specifically, they found three primary results. First, when trying to recognize a known melody, musicians experienced the FEP earlier than nonmusicians. Second, musicians experienced the IP later than nonmusicians. Finally, despite the fact that musicians experience the IP later than nonmusicians, they experience the RP earlier than nonmusicians. Our simulations explore the issues involved in capturing the pattern of these subjective states in a neural network framework. A preliminary version of this work appears in Vempala and Maida (2009).

To model the Dalla Bella et al. results we start with a core sequence recognition network (SRN) developed by Tank and Hopfield (1987). The neural network is a single-layer, time-evolving, winner-take-all (WTA) network. It has an associative input layer equipped with temporal delay filters intended to give the network robustness to distortions in the input sequence, such as incorrect items and time warping. This SRN was initially designed as a framework for spoken-word recognition where the sequence items were phonemes selected from the alphabet of phonemes for a particular language.

There are similarities between spoken-word recognition and isolated melody recognition. Dalla Bella et al. adapted an experimental paradigm known as the gating paradigm from the spoken-word recognition literature. In the melody recognition context, this involved incrementally presenting a melody to a listener and asking the listener to make a judgment after each presentation (e.g., Is this melody familiar?). More specifically, a listener was presented with the first note of a melody and asked to make a judgment, then the first two notes, and so on for the remaining notes of the melody. Dalla Bella et al. interpreted their results using the framework of cohort theory, which comes from the spoken-word recognition literature. The SRN was selected for its compatibility with cohort theory. It can be viewed as a reference implementation of cohort theory.

The present research is novel in at least the following ways. It introduces a reference implementation of cohort theory to the literature on melody recognition. As such, it provides a framework for interpreting and modeling the Dalla Bella et al. results. Our work is the first simulation study of these phenomena. We extend the Tank and Hopfield SRN with meta-level readout processes that monitor the dynamically evolving state of the network. This is used to model the subjective states of familiarity and confidence.

The original Tank and Hopfield SRN was intended to perform only sequence recognition tasks. It was previously not known whether this core network contained enough dynamic information to allow readout of subjective familiarity states, as observed by Dalla Bella et al. By adding the meta-level readout units and tuning them to give proper readouts, we have demonstrated, that this is indeed the case.

\(^2\)A melody is viewed as a temporally spaced sequence of notes. Recognition involves determining whether an input sequence of notes matches that of a stored melody in memory.
1.1. Melody Recognition and Spoken-word Recognition

Music shares features with language such as hierarchical and temporal structure, vocabulary, and tonal properties (Limb, 2006; McMullen and Saffran, 2004; Patel, 2008). Similarities between melody recognition and spoken-word recognition include the following. Both tasks require temporal working memory to consolidate input into a higher-order percept, as well as to compare it with information previously stored in long-term memory (LTM). Spoken works in a given language are made up from a limited inventory of speech sounds and tend to be phonologically similar to each other (Norris and McQueen, 2008). Many spoken words begin and end the same way as other words. Some words have other words embedded in them. This applies to melodies as well. Some melodies begin and end the same way as other melodies. Some melodies have other melodies embedded in them.

There are also differences between spoken words and melodies. Segmentation of spoken words by listeners is a recognized problem (Norris and McQueen, 2008). Although the speech signal consists of word-boundary cues to which listeners are sensitive, these cues are not completely reliable. Speakers do not segment their utterances for listeners, the way words are segmented with clear spaces in written sentences. This problem does not apply to melodies when melodic input is presented as consisting of individual pitches. Pitches are discrete units with clear onsets. They provide reliable cues for segmentation by the listener.

As a starting point, musical notes may be considered functionally analogous to phonemes in spoken words. A significant body of literature comprising experimental and computational work exists in the area of spoken-word recognition. One influential model of spoken-word recognition is the cohort model (Marslen-Wilson, 1987).

1.2. Tenets of Cohort Theory

Marslen-Wilson proposed the cohort model, in part, based on results of gating tasks that show spoken words can be reliably identified before complete acoustic-phonetic information is available (Grosjean, 1980; Cotton and Grosjean, 1984). In a spoken-word gating task, a listener is presented with increasing fractional increments of a word and is asked to identify the word at each increment. The isolation point, measured in milliseconds from word onset, is the point where the word is guessed correctly. The recognition point is the point where the listener is confident in this identification. This usually occurs after the isolation point because the listener needs additional evidence to reach a confidence threshold. In terms of the cohort model, the isolation point was viewed as the point at which the word could be discriminated from other members of the word-initial cohort.

In gating tasks, listeners identify words presented in sentential contexts about 100 milliseconds earlier than when presented in isolation. This implies that the recognition system uses top-down context in addition to sensory information. Evidence from gating tasks led Marslen-Wilson to propose the cohort model with the following features: multiple access, multiple assessment, and real-time efficiency. Multiple access refers to access of multiple candidates by mapping sensory input into the mental lexicon. The initial cohort of accessed candidates becomes available for selection. Multiple assessment refers to assessment of each candidate for selection based on semantic and syntactic appropriateness. This is where context affects selection. In addition, Marslen-Wilson found that more frequently occurring words have a temporary early advantage in the recognition process. His studies also suggested that top-down context does not inhibit bottom-up information.

To satisfy these requirements the recognition system should (1) be a distributed, parallel processing system rather than sequential to allow multiple access and assessment; (2) consist of
separate, computationally active recognition units, where each unit corresponds to a candidate in
the mental lexicon; (3) allow each unit to be activated with only bottom-up input and disallow
pre-selection based on top-down context; (4) enable recognition based on both accumulation
of positive evidence for a word and evidence that competitor words are not present; and (5) have
a tolerance for noise in the acoustic-phonetic input.

1.3. Effect of Training on Melody Recognition

Based on the analogy to spoken-word recognition, Dalla Bella et al. (2003) used the cohort
model and gating procedures to experimentally study melody recognition in musicians and non-
musicians. They presented gated melodies selected from a repertoire of French traditional songs
(Berthier, 1979). The melodies were divided into familiar and unfamiliar based on previously
collected norms (Peretz, Babai, Lussier, Hébert, and Gagnon, 1995).

Dalla Bella et al. (2003) conducted two experiments. The first experiment studied the emer-
gence of familiarity while listening to a melody. The second experiment studied melody isolation
and recognition.

In the first experiment, participants judged whether a melody was familiar after each incre-
ment of a gated presentation. This familiarity emergence point (FEP) was the note number in the
melody where the listener first judged the melody as familiar, and consistently maintained this
response for the rest of the trial. The FEP was measured only for familiar melodies.

The FEP occurred earlier in musicians than in nonmusicians. The FEP was interpreted as a
measure of the feeling of knowing at that point in the melody. Dalla Bella et al. (2003) interpreted
this finding based on the proposal of Koriat and Levy-Sadot (2001) that the feeling of knowing
is in part based on the total amount of accessed information in LTM. The trained musician, it is
hypothesized, will have more melodies stored in LTM than the nonmusician. Thus the musician
will access a larger initial cohort of melodies, leading to a stronger feeling of knowing, and an
earlier FEP.

Thus the FEP note number could be considered as a point in the time-course of melody
recognition at which the total amount of accessed information in LTM exceeds a threshold in
the listener for creating the perceived sense of familiarity. Linking this to a neural network,
this could be measured in an SRN as a dynamic state within the network where the amount of
accessed LTM information exceeds a threshold.

Based on the above premise, we extend the Dalla Bella et al. (2003) explanation by proposing
that the initial cohort of accessed melodies activates meta-level neurons correlated with a feeling
of familiarity. Such a familiarity set would need to be activated above a certain threshold to
signal a sense of familiarity. If the musician accesses a larger initial cohort than a nonmusician,
then the total activations of the neurons representing the initial cohort, serving as input to the
familiarity neurons, would be larger for the musician. This would drive the familiarity neurons
above threshold earlier for the musician.

In their second experiment, Dalla Bella et al. studied the time-course of recognition. Par-
ticipants sang the melody they thought was being presented and indicated their confidence level
during its gated presentation. Dalla Bella et al. established two recognition points. The isolation
point (IP) was defined as the point at which the participant guessed the melody. This was the note
number at which the participant correctly sang the next three notes of the melody and maintained
this response for the rest of the trial. The recognition point (RP) was defined as the point at which
the participant was completely confident in his or her judgment. This was the note number at
which the participant not only sang the melody correctly, but also indicated a maximum confi-
dence rating of seven. On average, the IP occurred earlier in nonmusicians than in musicians but
the RP occurred earlier in musicians than in nonmusicians. Dalla Bella et al. proposed that since musicians access a larger initial cohort than nonmusicians, they must examine a larger number of candidates before isolating the correct melody. This accounts for the earlier occurrence of the IP in nonmusicians.

They also proposed that after isolating a melody, on average, musicians are more confident in their judgments than nonmusicians, resulting in a slightly earlier RP than nonmusicians. Dalla Bella et al. viewed their listener’s maximum confidence rating of seven to be a reflection of their sense of maximum confidence after isolating the melody. Thus, the RP was the point where the participant was both accurate and confident in recognition of the melody. Experimental studies showing dissociations between confidence and accuracy suggest the role of other factors in determining confidence level, besides strength of the memory trace (Busey, Turnicilff, Loftus, and Loftus, 2000; Chua, Schacter, Rand-Giovannetti, and Sperling, 2006). In a task where subjects chose an answer to a two-alternative question, Koriat (2008) found that a higher level of familiarity about the question’s domain could increase the subject’s confidence level independently of response accuracy. One method of finding the RP in an SRN is by using a meta-level process which produces this notion of confidence depending on the level of familiarity, by monitoring the state of the network. A novel feature of our simulation is that the meta-level readout unit that signals the FEP is used as an input to the unit that signals the RP. Thus, our model is able to assign a causal role to the sense of familiarity. Among other possible roles, it helps in determining the RP.

The pattern of results in Dalla Bella et al. (2003) are shown in Figure 1. We developed a simulation to model the hypothesized cognitive processes underlying the FEP, IP, and RP with the intention of capturing the qualitative configuration of these six points.

Figure 1: Time course of melody recognition in musicians and nonmusicians (adapted from Dalla Bella et al. (2003)).

1.4. Motivation for using the SRN

Several influential computational models of spoken-word recognition exist such as TRACE (McClelland and Elman, 1986), Shortlist (Norris, 1994), Shortlist B (Norris and McQueen, 2008), Merge (Norris et al., 2000), NAM (Luce and Pisoni, 1998) and PARSYN (Auer and Luce, 2005). Both TRACE and Shortlist are activation-based, parallel processing models that satisfy many requirements of the cohort model such as multiple access and selection.

TRACE allows integration of top-down context. It consists of three levels of processing. These are the feature level, phoneme level, and word level. Lateral inhibition within each level allows for competition. Feedforward and feedback excitatory connections between processing
levels allow facilitation of unit activations. TRACE compensates for noisy input by rendering the input as a bank of features that could be fed into the units in the feature level. Connections in TRACE are bi-directional allowing top-down context to influence bottom-up input without restrictions. This is where TRACE differs from the cohort model which places restrictions on the influence of top-down context. In the cohort model, top-down context can only have limited facilitory effects but not inhibitory effects on candidates in the cohort.

Shortlist, designed by Norris (1994), is an alternative to TRACE. It was intended to achieve the functionality proposed by the cohort model. Shortlist has a simpler architecture than TRACE but shares functional similarities. It has two levels of processing. In the first level, candidate words are associated with each phoneme. A lexical search procedure is performed for each phonemic input, during which scores are assigned to each candidate word for the input phoneme. These scores are constantly updated with the presentation of every new phoneme. Candidates are assigned negative scores based on the amount of mismatch. The lexical search procedure determines the membership of candidates in the candidate set. The candidate set also known as “the shortlist” is similar to the word-initial cohort in the cohort model. At the second level, the selected candidates are connected to a lexical network in which candidates with overlapping input compete through mutual inhibitory links. The main difference between Shortlist and TRACE is the entirely bottom-up functionality of Shortlist. Top-down influences, as used in TRACE, are not incorporated.

From a design standpoint Shortlist is simpler than TRACE because of the direct link between phonemes and words. TRACE offers a functional hierarchy by separating the perception-recognition process into three distinct levels consisting of features, phonemes, and words.

The Merge model is similar in architecture to Shortlist with the exception of one additional component. It is based on the assumption that feedback from lexical units to prelexical units as in the TRACE model are unnecessary. However, it recognizes that lexical information can sometimes help in identifying phonemes. So, a phoneme decision component is added as a third part of the model. The phoneme decision component receives inputs from the prelexical units as well as the lexical units.

Norris and McQueen (2008) designed a Bayesian version of Shortlist known as Shortlist B. It uses a feedforward architecture, no online feedback, and evaluates multiple lexical hypotheses, similar to Shortlist. However, Shortlist B is entirely Bayesian and not activation-based. It assumes that listeners make optimal Bayesian decisions during spoken-word recognition. Unlike in Shortlist where the input is a sequence of discrete phonemes, the input to Shortlist B is a sequence of multiple phoneme probabilities over three time slices per segment derived from listeners’ performance in a gating study. The goal of Shortlist B is to find the posterior probability of the word given the evidence, \( P(\text{Word}_i|\text{Evidence}) \). Using Bayes’ theorem, \( P(\text{Word}_i|\text{Evidence}) \) is proportional to \( P(\text{Evidence}|\text{Word}_i)P(\text{Word}_i) \). Shortlist B obtains prior probabilities of words \( P(\text{Word}_i) \) from the frequency of occurrence of words in the CELEX database (Baayen, Piepenbrock, and Gulikers, 1994).

Shortlist B assumes the evidence consists of sublexical units which are phonemes. Therefore \( P(\text{Evidence}|\text{Word}_i) \) needs to be estimated using phoneme probabilities which serve as input to the model over three time slices. Shortlist B uses perceptual confusion data to directly estimate \( P(\text{Phoneme}_i|\text{Evidence}) \) and then find \( P(\text{Evidence}|\text{Word}_i) \) which is the product of the probabilities of all the phonemes in the \( \text{Word}_i \). In continuous speech, since a word can overlap with several other words, Shortlist B calculates path probabilities to find out if a word falls on a high or a low probability path, and incorporates path probabilities within its measure of word probability. The probability of each path \( P(\text{Path}_i|\text{Evidence}) \) is computed by normalizing over
the sum of all path likelihoods. The likelihood of each path \( P(Evidence | Path_i) \) is computed as the product of all terms \( P(Evidence | Word_j)P(Word_j) \) for each word in the path \( (j = 1 \) to \( w \), where \( w \) is the number of words in the path). Now, \( P(Word_j | Evidence) \) is calculated by summing all the \( P(Path_i | Evidence) \) terms for the number of paths the word falls on.

Tank and Hopfield (1987) designed a neural network to recognize sequences unfolding in time, such as phoneme sequences in spoken words. Its architecture is compatible with cohort theory. Their sequence recognition neural network (SRN) recognized noisy and time-warped sequences. The TRACE model simulates short-term memory by duplicating units over discrete time slices. The Tank and Hopfield SRN uses a more elegant method of simulating short-term memory. It uses delay functions to account for short-term storage of previous notes in a sequence through delay filters. Despite the existence of many computational models, the Tank and Hopfield SRN is more appropriate for modeling the Dalla Bella et al. study because it satisfies five key requirements of the cohort model using a simple single layer architecture.

These are (1) The SRN has a parallel processing architecture enabling multiple candidates to be accessed and assessed based on input. (2) The SRN satisfies the cohort model’s LTM representational specificity requirement by assigning separate, computationally active sequence recognition (SR) units to specific temporal sequences stored in LTM. (3) The SRN is an activation-based network where activation levels of matching candidates are represented by outputs of SR units. In addition, each SR unit is activated based on bottom-up sensory input and not top-down contextual feedback, as specified in the cohort model. (4) The cohort model shows tolerance to noise within the sensory input signal analogous to the human recognition system’s ability. The SRN is designed so that distorted versions of sequences can still activate the corresponding SR units. These distortions may either be in the form of time-warp or erroneous symbols. (5) The SRN uses a time-evolving winner-take-all (WTA) mechanism among SR units, satisfying the cohort model’s requirement that each unit should take into account the behavior of competing units. The WTA mechanism is implemented through inhibitory feedback connections between competing units.

The Tank and Hopfield SRN is arguably the simplest single layer architecture that achieves these requirements. One of our goals is to obtain a minimal model consistent with these requirements. The cohort model does not address acquisition of new memories. It examines spoken-word recognition processes under the assumption that acquisition has already occurred. Dalla Bella et al. also look at melody recognition processes in terms of the cohort model without focusing on melody acquisition. Since both Dalla Bella et al. and the cohort model focus on recognition based on existing LTM states with no learning, the SRN is suitable for modeling melody recognition because weights are preset on the basis of LTM contents without learning. This enables comparisons reflecting the corpuses of a musician and a nonmusician. Existing computational models in spoken-word recognition only model recognition of representations in LTM. They do not model familiarity emergence. The Tank and Hopfield SRN, because of its simple architecture, may be extended for modeling familiarity emergence.

2. Modeling Framework

Our core network is adapted from the Tank and Hopfield SRN and is implemented in the JAVA programming language. This core network models the IP. It consists of a recurrent inhibitory network of sequence recognition (SR) units. There is one SR unit for each stored melody. The inhibitory connections implement a time-evolving, winner-take-all (WTA) competition among
the SR units. Inputs to the network are a set of pitch detectors spanning two octaves in half-step (semitone) increments. Each detector is connected to a bank of delay filters that act as an acoustic memory. The core network is extended with two meta-level readout units. One unit, called the familiarity unit, models the FEP by monitoring SR unit activations in the core network. A second unit, called the recognition unit, uses inputs from the core network and the familiarity unit to model the RP.

2.1. Corpus Selection
To distinguish the musician’s LTM from the nonmusician’s, the musician’s corpus is six times larger than the nonmusician’s. The nonmusician’s corpus consists of five melodies and the musician’s corpus has an additional 25. The five melodies common to both corpuses are intended to be familiar, similar to the familiar melodies used in Dalla Bella et al. The full corpus is listed in Table 1 of the Appendix. The five popular melodies were selected on the basis of popularity from best-selling artists listed in Recording Industry Association of America (2008) and Billboard.com (2008). The musician network was intended to represent a musician proficient in a specific musical genre. The selected genre was jazz. Since jazz musicians study several jazz standards, we selected the remaining 25 melodies from jazz standards (Schoenberg, 2002; Hal Leonard Corporation, 2002; Jazzstandards.com, 2009).

Each melody was a small subset of the song from which it was selected. It lasted 10-25 quarter-note time units, capturing an essential part of the song (e.g. chorus, beginning few notes, bass line).

2.2. Network Inputs and Pitch Detectors
For recognition tests, a melody was input to the network as a sequence of notes. All melody sequences fell within a two-octave range. Each octave had 12 notes: C, C#, D, D#, E, F, F#, G, G#, A, A#, B. Since melodies were limited to 24 pitch tones, networks had 24 pitch detector units. Pitch detectors for tones in the first octave were denoted by the note name followed by the suffix 1 (e.g. C1, C#1, D1). Detectors for the second octave were indicated by the suffix 2 (e.g. C2, C#2, D2). Table 2 in the appendix shows the representations of the 30-melody corpus using this symbolic notation.

Detectors gave binary outputs of 0 or 1 depending on the absence or presence of the pitch tone at a particular time step. Following Tank and Hopfield the notation, \( D_X(t) \), indicates the presence or absence of feature \( X \) at time \( t \). In our simulations, a feature is a note. Specifically,

\[
D_X(t) = \begin{cases} 
1 & \text{if note } X \text{ is present at } t \\
0 & \text{otherwise}
\end{cases}
\]

Rests were denoted with the letter R. Quarter, half, and whole note durations were represented as prefixes to the note name. The representation did not capture rhythmic nuances based on note onsets, such as the difference between one half and two quarter notes. Duration of a sequence was computed as a sum of quarter note steps. Although key invariance while listening to melodies and key detection of a melody are important aspects of melody recognition, these issues were beyond the scope of the current research and all melodies were transposed to the key of C. Each pitch detector is connected to a bank of delay filters as shown in Figure 2.
2.3. Delay Filters

The delay filters (Tank and Hopfield, 1987) provide a model of acoustic memory. Design of these filters was influenced by the biophysics of down-chirp pulse detection neurons in birds such as the barn owl (Hopfield and Tank, 1989). Hopfield and Tank outlined a possible mechanism for down-chirp pulse detection where high frequencies occurring earlier in the pulse propagate along low velocity axons, and low frequencies occurring later propagate along high velocity axons, thereby arriving at the detection neuron simultaneously. Outputs of the delay filters are specified by continuous delay functions shown in Figure 3. The delay functions enable the peaks of pitch tones within a melody to synchronously reach their maximum at the end of the input sequence. The functions take the form

\[ f_k(t) = \left( \frac{t}{k} \right)^n e^{n(1 - \frac{t}{k})}, \]

where \( k \) is the number of time steps remaining for the sequence to complete and determines the width of the curve, \( t \) is the time elapsed since the curve was initiated, and \( n \) is a constant. A delay curve \( f_k(t) \) peaks at time value \( t = k \). Tones occurring closer to sequence completion have shorter durations (\( k \)'s) and peak earlier than tones occurring nearer the beginning of the sequence. \( n \) determines the position of the peak with respect to sequence duration \( \tau \) for a given \( k \). In Figure 3, if curve \( k = 1 \) corresponds to the last note in the melody, curve \( k = 4 \) corresponds to a note beginning three steps from the end of the melody, and curve \( k = 7 \) corresponds to a note beginning six steps from the end of the melody, then all three curves will peak at the same time at the end of the melody. In Figure 2, the parameter \( k \) is set to \( \tau + 1 - t \), where \( \tau \) is the length of the melody and \( t \) is the time of onset of the note. For our simulations, \( n \) was set to 5.

2.4. SR Units and Connection Weights

SR units act as melody detectors. If a network stores a corpus of 30 melodies then there will be 30 SR units. Each SR unit has a membrane potential \( u \) and an output voltage, \( V \), as shown in Figure 4. \( V \) varies between 0 and 1 and is also known as the unit’s activation. SR unit \( i \) is a leaky integrator that sends membrane potential, \( u_i \), through a (nonlinear) logistic-sigmoid function. The activation, \( V_i \), for a unit \( i \), is a logistic sigmoid function of the membrane potential, \( u_i \),

\[ V_i = g(u_i) = \frac{1}{1 + e^{-\frac{2}{9} u_i}}. \]
Figure 3: Delay curves, $f_k(t)$. Values of $k$ are 1, 4, and 7. Notice that the corresponding curve peaks at $t = k$.

Figure 4: The structure of an SR unit consists of a leaky integrator of membrane potential $u$ with nonlinear output $g(u) = V$.

The output of each delay function is sent to the input of each SR unit. The connection from the delay function to the SR unit is mediated by a weight which is preset to either 0 or 1 and does not undergo learning. The value of the weight indicates absence or presence of a connection. Following (Tank and Hopfield, 1987), connection weights are denoted $T_{i,X;k}$, where $i$ denotes the SR unit that receives the connection, $X$ denotes the note value that was detected, and $k$ denotes the particular delay filter. The set of possible weights entering SR unit 1 are shown in Figure 2.

Connection weights are preset according to a Hebbian formula adapted to sequence representations as follows. Assume SR unit $i$ recognizes some melody sequence, say, C1, E1, G1, C1. Each tone is one time unit long and the duration of the melody, $\tau$, is 4. When the melody is input to the network, pitch detector C1 has an output of 1 in the first time unit, detector E1 has an output of 1 in the first time unit, and so forth. Since C1 is the first item in the four-note sequence, $k = \tau + 1 - t = 5 - 1 = 4$. Therefore, $T_{i,C1;4} = 1$. By similar reasoning, $T_{i,E1;3} = T_{i,G1;2} = T_{i,C1;1} = 1$. All other weights coming into unit $i$ are zero.

SR units are connected to each other by inhibitory weights of strength $\alpha$ as shown in Figure 5. These connections form a WTA network. SR units do not connect to themselves. The update
equation for an SR unit, $i$, is given in the equation below.

$$\frac{du_i}{dt} = \frac{-u_i}{R} - \alpha \sum_{j \neq i} V_j - \gamma + \beta \sum_{k} \int_{0}^{\infty} T_{i,X,k} f_k(t') D_X(t-t') \, dt'$$

(2)

The membrane resistance, $R$, was set to 0.5. The first term on the right-hand side of the equation $-u_i/R$ models the leaky integrator. The second term models the WTA network with inhibitory connection weight values of $-\alpha$ ($\alpha = 2.5$). The third term, $-\gamma$, is a global inhibition set to $\gamma = 2.5$. Its purpose is to prevent false positives. The fourth term provides the excitatory input. It sums a convolution integral over all of the input notes and all of the time delays. It contributes to an increase in the input voltage $u_i$ when input matches the unit’s connection weight pattern. The product of the pitch detector output with the delay function is gated by the connection weights, $T_{i,X,k}$. $\beta$ was set to 0.6 and scales the excitatory inputs for the summation integral term in the circuit dynamics equation. This equation was integrated using the forward-Euler method using a step size of 0.1.

2.5. Modeling the IP using the Core SRN

As explained in Section 1.3, the IP was measured by Dalla Bella et al. as the note number at which the participant correctly sang the next three notes of the melody beyond the presented gate. Accordingly, we defined the IP for our network as follows. It was the time step at which the activation of some SR unit (a) reached 0.7 and was greater than the activations of all other competitor SR units in the network, and (b) continued to remain higher than its competitors for the next three consecutive time units, analogous to a participant correctly guessing the next three notes of the melody (or until sequence completion). After the start of the input melody, output $V_i$'s of all SR units were compared to find the unit with the highest activation at each time step. If an SR unit’s $V_i$ was at least 0.7, the time step, $IP_t$, was noted. If the SR unit’s activation remained higher than the $V_i$’s of the competitor SR units for the next three steps, then $IP_t$ was taken as the time of the IP.

2.6. Modeling the FEP using a Meta-level Familiarity Unit

Dalla Bella et al.’s explanation for the cause of the earlier FEP in musicians was that musicians access a larger initial cohort of melodies than nonmusicians because of a larger corpus being stored in LTM. This gives a stronger ‘feeling of knowing’ and earlier FEP as explained in Section 1.3. We interpret Dalla Bella et al.’s explanation by proposing that the initial cohort of
accessed melodies activates familiarity neurons correlated with a feeling of familiarity. We represent these neurons by a single familiarity unit. This unit needs to be activated above a threshold to create a sense of familiarity. If the musician accesses a larger initial cohort than the nonmusician, then the total activations of the units representing the initial cohort, serving as input to the familiarity unit, would be larger for the musician. This would drive the familiarity unit above threshold earlier for the musician. We extended the core network by adding a familiarity unit. The feeling of familiarity was evoked by the activity of this unit. The familiarity unit, \( F \), has a logistic sigmoid activation function which takes as input the sum of activations of the SR units in the network. It is described by the formula below.

\[
F(t) = \frac{1}{1 + e^{-\sum_i V_i(t)}}
\]

The familiarity unit threshold for signaling occurrence of the FEP was 0.65. The time of the FEP, \( FEP_t \), is the time at which \( F(t) \) goes above threshold. Two criteria influenced this choice: (1) Since the FEP occurs earlier than the IP and RP, and all occur prior to melody completion, the FEP must occur earlier in the melody ahead of sequence completion, and (2) the FEP should reach threshold when there is increase in the activity of the SR units after input starts, indicating perception of melodic input.

2.7. Modeling the RP using a Meta-level Recognition Unit

Dalla Bella et al. measured the RP by the note number at which the listener (a) sang the melody correctly, and (b) provided a maximum confidence (7-point scale) for the first of three consecutive presentations. Thus, the RP should take both factors, confidence and accuracy, into consideration. Dissociations between confidence and accuracy in experimental studies suggest the role of other factors in determining the confidence level, besides strength of the memory trace alone. As explained in Section 1.3, Koriat (2008) found that a higher level of familiarity about the question’s domain could increase the subject’s confidence level, independently of response accuracy. Based on the hypotheses that greater accuracy and familiarity both cause higher confidence, we assumed that a meta-level recognition process computes the recognition point by monitoring (1) the WTA network used in finding the IP, as well as (2) the familiarity unit for measuring confidence. To take into account that a sense of familiarity can influence confidence, we used the formula below to determine the time of the RP.

\[
RP_t = IP_t + 2(l_i - IP_t)(1 - F(IP_t))
\]

The second term computes a cost in time units based on strength of the familiarity unit output and the amount of evidence already presented to the network in reaching the IP. The symbol \( l_i \) is the length of the melody presented to the network in time units. The cost is added to the IP to compute the RP. Further, the greater the strength of the familiarity output at the time of the isolation point, \( F(IP_t) \), the lower the cost. Similarly, the greater the amount of evidence already presented, the lower the cost.

3. Basic Simulations, Results, and Discussion

A network was constructed with five stored melodies to represent a nonmusician. A second network with 30 stored melodies was constructed to represent a musician. For all simulations,
Figure 6: SR activations in core network when tested with familiar melody 2. The test melody begins at step 3. Melody 2 is 15 time steps long and ends with the start of step 19. In both musician and nonmusician networks, SR unit 2 is activated. (A) Nonmusician network with corpus of five melodies. (B) Musician network with corpus of 30 melodies.

the input voltages, $u_i$, for the SR units was initialized to $-5$ so that the output voltages, $V_i$, were approximately 0 at $t = 0$. To allow the network to stabilize, test melodies were presented beginning at time step 3. The five familiar melodies (melodies one through five) were presented as test melodies to both the musician and nonmusician networks.

Figure 6 shows the results when test melody 2 was presented to the networks. The melody-specific SR unit corresponding to melody 2 (the presented melody) responded most strongly with activation levels decisively above those of their competitor units in both musician and nonmusician networks. Figure 7 shows the results when test melody 3 was presented to the networks. Again, the SR unit representing the presented melody responded most strongly with activation levels decisively above those of their competitor units in both musician and nonmusician networks.

3.1. FEP, IP, and RP Results

Familiarity unit activations for the simulations in Figures 6 and 7 are shown in Figure 8. In both cases, the musician familiarity unit reaches the activation threshold of 0.65 for the FEP earlier than the nonmusician unit. We computed mean values for FEP$_t$ over the five test melodies for both musician and nonmusician. For the nonmusician, the mean value of the FEP$_t$ was 11.0 with standard error 0.55. For the musician, the mean FEP$_t$ was 6.4 with a standard error of 0.24. These values were computed relative to test melody onset time so 3 was subtracted from all of the values.

For the nonmusician network, the mean IP$_t$ was 12.4 time steps after the onset of the test melody (standard error of 0.68). The mean IP$_t$ for the musician network was 13.2 time steps into the melody (standard error of 0.66). The nonmusician IP was 0.8 time units earlier than the musician network. These results were comparable to Dalla Bella et al.’s results where the nonmusicians IP occurred 0.3 to 0.4 notes earlier than the musician.

For the nonmusician network, the mean RP$_t$ was 15.0 time steps after the onset of the test melody (standard error of 0.79). The mean RP$_t$ for the musician network was 14.9 time steps
into the melody (standard error of 0.81). The nonmusician IP was 0.8 time units earlier than the musician network. These results indicated that the musician required less information than the nonmusician to reach the RP after the IP, despite reaching the IP 0.6 time units later, because of a higher level of confidence indicated by the level of familiarity. Figure 9 shows the final pattern of results for the FEP, IP, and RP. Part A shows the Dalla Bella et al. results and part B shows our results. Our simulation results captured the rank ordering of the FEP, the IP, and the RP, and qualitatively matched Dalla Bella et al.’s results.

### 3.2. Melodic Similarity Measurements

Müllensiefen and Frieler (2004) provide evidence that edit distance measurements that use a rich symbolic representation compare well with human melody similarity judgments. Edit distance is defined as the minimum number of edits required to transform one string into another using insertion, deletion, and substitution of single characters. We compared a symbolic representation of melody similarity with a network measure of melody similarity. The network measure was maximum activation of an SR unit to some test melody. Since the SR unit represents a stored melody, its maximum activation is a measure of how close it thinks the test melody is to the stored melody, taking into account competition from other units.

We started by computing edit distances (Wagner and Fischer, 1974) between a test melody, such as melody 2, and each of the 30 stored melodies in the musician corpus. We then normalized these edit distances by the length of the melodies. Specifically, for two melodies \( m_1 \) and \( m_2 \), we divided their edit distance by the length of the longest of \( m_1 \) or \( m_2 \). This gave a normalized edit distance with values ranging from zero to one. Maximum activation of an SR unit in the network in response to a test melody is a measure of network similarity of the stored melody to the test melody. To compare network similarity to symbolic melody similarity, we converted the normalized edit distance measure to a similarity measure by subtracting it from 1. We called this symbolic melody similarity. We found that network similarity correlated with symbolic melody similarity.
Figure 8: Familiarity unit activation. Melody presentation begins at step 3. FEP threshold is 0.65. Dots indicate emergence of FEP for musician (corpus 30) and nonmusician (corpus 5). (A) Test melody 2. FEP<sub>t</sub> = 10.1 for musician and 14.1 for nonmusician. (B) Test melody 3. FEP<sub>t</sub> = 10.9 for musician and 14.5 for nonmusician.
Figure 9: Comparison of Dalla Bella et al. results with the SRN simulation results. (A) The Dalla Bella et al. results repeated from Figure 1. (B) The simulation results of this section.

Figure 10 shows the results for melody 2 and melody 3. Part A shows the results for melody 2 and part B shows the results for melody 3. The dashed line shows symbolic similarity of the test melodies to each of the stored melodies. Since test melody 2 is the same as stored melody 2, it is maximally similar to itself. This is reflected by the peak in the graph of part A. For melody 2, its symbolic similarity to other stored melodies is virtually zero. The same pattern is predominantly true for melody 3 in part B. The solid lines show the maximum activations of the SR units that correspond to the stored melodies. Specifically, the solid line in part A shows the maximum activation of each SR unit when melody 2 is given as the test melody to the network. Finally, the two measures of similarity are correlated. The correlation between symbolic similarity and network similarity when melody 2 is used as the test melody is \( r = 0.81 \). When melody 3 is used, the correlation is \( r = 0.86 \).

3.3. Summary of Model Parameters and Their Causal Interpretations

When interpreting these data, a question arises as to the number of model parameters, their associated degrees of freedom, and the explanatory power of the model. Here we summarize and review the number of free parameters, thresholds, and neural weights. First, we note that the neural weights do not contribute degrees of freedom to the model. The model has no adjustable connection weights. The weights in the core SRN are binary (value of 0 or 1) and are preset according to a Hebbian formula to store the corpus of melodies. For the familiarity unit, all of the incoming weights have a preset fixed value of one. Finally, the RP unit does not have incoming weights because it computes an arithmetic function of its inputs.
The model has five free parameters that are nonthreshold parameters. One of these parameters, \( n \), is used in Formula 1. This parameter controls the width of the \( f_k \) temporal filters. It affects whether they are broadly or narrowly tuned in time. As can be seen in Figure 3, the value \( n = 5 \) was chosen to give the filters moderately broad temporal tuning. The remaining parameters appear in Equation 2 and control the operational characteristics of the core SRN. For instance, the parameters \( \alpha = 2.5 \) and \( \beta = 0.6 \) control the balance of excitation and feedback inhibition so that the network can act as a dynamic memory. The parameter \( \gamma = 2.5 \) controls the global inhibition and is needed to prevent false positives. The parameter \( R = 0.5 \) affects the network’s activation time constant.

Lastly, there are two threshold parameters. The first is the FEP threshold which is set to 0.65. From Figure 8, it can be seen that activation of the familiarity unit is always higher for the musician corpus than the nonmusician corpus. Thus, the fact that the simulation of the FEP occurs earlier for musicians is robust to changes in threshold and the exact threshold value is not critical. The other threshold parameter is takes value 0.7 for the IP threshold. By examining Figures 6 and 7, a similar argument can be made for the robustness of the IP threshold.

4. Discussion and Conclusion

This paper has shown how to use a sequence recognition neural network to model human melody familiarity and recognition judgments between musicians and nonmusicians. Cohort theory from spoken word recognition was used to link the experiments of Dalla Bella et al. (2003) with the sequence recognition network architecture of Tank and Hopfield (1987). The core sequence recognition network was viewed as a minimal representation of the tenets of cohort theory. Using a network augmented with meta-level readout units, we successfully captured the qualitative pattern of events in the Dalla Bella et al. study. Specifically, we showed how a simulation could model subjective mental states known to occur in melody recognition, namely, the familiarity emergence point, the isolation point, and the recognition point. Further, we were able to assign a causal function to one of the readout units. In particular, the output of the familiarity unit was used as an input to the recognition unit. Since our model attributes a causal role to the familiarity unit, it does not view the FEP as an epiphenomenon.

Let us make this argument more explicit. It is not in dispute that humans have a subjective sense of familiarity as they perform the melody gating task. Further, it is axiomatic that this subjective mental state will have some kind of neural correlate. What is the form of this neural correlate? Plausibly, it may take the form of activity within some population of neurons. Since activity in this form can be read out by downstream neurons, the neural activity representing the sense of familiarity can have a causal impact on other parts of the system.

As stated earlier, the original Tank and Hopfield SRN was intended to perform only sequence recognition tasks. It was previously not known whether this core network contained enough dynamic information to allow readout of subjective familiarity states. By adding the meta-level readout units and tuning them to give proper readouts, we have demonstrated by existence proof, that this information does reside in the core network. Since the meta-level units take only the dynamic state of the core network as input, we could not have done this if the information was not already contained in the core SRN. The readout units only made the subjective states symbolically explicit. A question for additional further research is to explain the functions that such subjective cognitive states may have. For instance, our model gives the sense of familiarity a role in influencing confidence. It is possible that confidence, in turn, plays a role in adaptively influencing other aspects of behavior. Our model does not address this.
One limitation of the present research is that we have modeled the difference between the nonmusician and the musician strictly in terms of corpus size. Corpus size, while being important, is unlikely to be the only factor that differentiates musicians from nonmusicians. The nonmusician, such as a music critic, could be an informed listener of specific genres with an equally large corpus as that of a musician. Also not captured is the musician’s increased level of involvement in music as compared to the nonmusician, both through training as well as performance. Increased involvement could manifest itself in the form of activating richer mental representations such as engagement of cognitive and motor schemas, and greater contextual feedback while listening to melodies. In the present research, we have ignored these possible factors in order to focus on the effect of corpus size. This was motivated by Dalla Bella et al.’s original interpretation in terms of cohort theory and total amount of accessed information.

The current model has other limitations that derive from its simplicity. All melodies have the same strength in their LTM representation. Inclusion of frequency effects, would allow us to associate higher prior probabilities to melodies that are heard more often. While the current model illustrates accessing an initial cohort in the form of activations based on pitch-related input and their comparison with LTM, it does not show the facilitatory effects of top-down melodic context based on syntactic information. To account for such effects, it will be necessary to add hierarchy to the model. Another limitation is concerned with acquisition of melodies. Although the network models the recognition of melodies by SR units for an existing LTM state, it does not learn novel melodies.

References


18
A. Melodies Used and Their Transcriptions

This appendix lists the melodies used in the simulation. It also includes their transcribed symbolic representations. Table 1 lists the melody names, their ID numbers, and their length in quarter note time steps. The first five melodies in Table 1 constitute the nonmusician’s corpus. The full set of 30 melodies constitute the musician’s corpus. Table 2 shows the transcribed symbolic representations. The notation is explained in Section 1 of this paper.
Figure 10: Comparison of symbolic melody similarity with maximum SR unit activation. (A) Similarity of melody 2 to other melodies in corpus compared with maximum SR unit activation ($r = 0.81$). (B) Similarity of melody 2 to other melodies in corpus compared with maximum SR unit activation ($r = 0.86$).
<table>
<thead>
<tr>
<th>Melody #</th>
<th>Name</th>
<th>Length</th>
<th>Melody #</th>
<th>Name</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Banana Boat Song</td>
<td>15</td>
<td>16.</td>
<td>The Man I Love</td>
<td>14</td>
</tr>
<tr>
<td>2.</td>
<td>Beat It</td>
<td>15</td>
<td>17.</td>
<td>West End Blues</td>
<td>25</td>
</tr>
<tr>
<td>3.</td>
<td>Hound Dog</td>
<td>16</td>
<td>18.</td>
<td>Cottontail</td>
<td>16</td>
</tr>
<tr>
<td>4.</td>
<td>We Don’t Need No Education</td>
<td>16</td>
<td>19.</td>
<td>Reflections</td>
<td>19</td>
</tr>
<tr>
<td>6.</td>
<td>12th Street Rag</td>
<td>21</td>
<td>21.</td>
<td>All the Things You Are</td>
<td>15</td>
</tr>
<tr>
<td>7.</td>
<td>Autumn Leaves</td>
<td>18</td>
<td>22.</td>
<td>'Round Midnight</td>
<td>15</td>
</tr>
<tr>
<td>8.</td>
<td>Blue Horizon</td>
<td>17</td>
<td>23.</td>
<td>What is this Thing Called Love</td>
<td>11</td>
</tr>
<tr>
<td>9.</td>
<td>Boplicity</td>
<td>10</td>
<td>24.</td>
<td>Willow Weep For Me</td>
<td>17</td>
</tr>
<tr>
<td>10.</td>
<td>Haitian Fight Song</td>
<td>16</td>
<td>25.</td>
<td>On Green Dolphin Street</td>
<td>16</td>
</tr>
<tr>
<td>11.</td>
<td>I Got a Right to Sing the Blues</td>
<td>16</td>
<td>26.</td>
<td>Smoke Gets in Your Eyes</td>
<td>18</td>
</tr>
<tr>
<td>12.</td>
<td>Lullaby of Birdland</td>
<td>20</td>
<td>27.</td>
<td>S’Wonderful</td>
<td>19</td>
</tr>
<tr>
<td>15.</td>
<td>Well You Needn’t</td>
<td>16</td>
<td>30.</td>
<td>Darn That Dream</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 1: The corpus of 30 melodies used in the simulations. The table shows the melody ID number, the melody name, and the melody length in quarter note steps.
Table 2: The symbolic note representations used for melodies in the simulations.