MELODIC MEMORY AND ITS DEPENDENCE ON FAMILIARITY AND DIFFICULTY

Mariana E. Benassi-Werke
Psychobiology Dept.
UNIFESP, Brazil
marianawerke@yahoo.com.br

Marcelo Queiroz
Comp. Science Dept.
USP, Brazil
mqz@ime.usp.br

Nayana G. Germano
Psychobiology Dept.
UNIFESP, Brazil
nayanager@hotmail.com

Maria Gabriela M. Oliveira
Psychobiology Dept.
UNIFESP, Brazil
mgabi@psicobio.epm.br

ABSTRACT

This paper addresses one aspect of human music cognition, which is the recollection of melodic sequences stored in short-term memory, and the manipulation of such items in working memory, by measuring spans of successfully recalled melodic sequences. In order to avoid long-term memory collaboration in this task, short-term memory measurements are made using randomly-generated melodic sequences, which in turn may sound difficult and unfamiliar to many experimental subjects. We investigate the dependence of melodic span measures on such aspects as familiarity and difficulty, both in direct-order recalling (as it relates to short-term memory capacity) and in inverse-order recalling (as it relates to working memory capacity). We also discuss the relation of these measurements to cognitive models of short-term and working memory for verbal and melodic material.

1. INTRODUCTION

Understanding human music cognition is a colossal task, which nevertheless must be undertaken. Besides its scientific interest per se, better understanding the way we humans process musical information should allow further developments in computational psychoacoustics, particularly in cognitive models for automatic feature extraction, with implications for both automatic musical analysis and computer-based sound synthesis.

This study is a small contribution to the understanding of one very restricted musical cognitive task, namely our ability to reproduce melodic fragments we never heard before. This ability involves a part of our cognition usually referred to as short-term memory, which has been extensively studied in the field of experimental psychology [1]. More recently, Baddeley and Hitch [2] proposed a refined model called *working memory*, that subsumed the notion of short-term memory, and eventually became the de facto standard model referring to short-term memory.

We wish to investigate the response level of our working memory to simple tasks such as reproducing a melodic fragment in direct order or in inverse order (also called reverse or retrograde, not to be confused with melodic inversion). Such an experiment is a straightforward transposition of classical experiments with working memory using sequences of digits or words, which in our case is aimed at identifying common or disparate elements in the processing of verbal and melodic information. As in the verbal case, random sequences should be used in order to avoid the contribution of long-term memory, which we routinely use in the memorization of whole musical pieces, for instance.

Random melodies can be quite hard to recall due to many concurrent factors, which might be empirically hypothesized, such as the number of distinct tones in a sequence, or the interval relations between adjacent notes. Similar factors might also affect the memorization of numbers or words, although interval relations may have no meaning in most non-musical contexts. Other factors, such as word-length and phonological similarity, are well-known to affect verbal memorization [3].

The difference in number and internal organization of distinct tones is also a characteristic feature of musical scales, such as the pentatonic (5-tone), diatonic (7-tone) and chromatic (12-tone) scales. Particularly in western musical education, diatonic and chromatic scales are everywhere present, from church modes through classical tonal music to 20th-century atonality. Yet the frequency with which diatonic scales have been employed in western folk, popular and classical music overshadows those relatively few pristine examples of entirely chromatic compositions. It is relatively safe to say that the average person growing up in western civilization is biased towards being more familiar with diatonic rather than chromatic musical examples.

Different interval relations between adjacent notes might also affect differently our perception, memorization and ability to reproduce melodic sequences. To name a few cases where such difference is mentioned, Fux's *Gradus ad Parnassum* of 1725 advises composers not to use large melodic leaps such as sixths or sevenths because they are hard to sing, and Nicola Vaccaj's *Metodo Pratico di Canto* of 1832 is arranged progressively according to melodic leaps. This suggests that smaller intervals (seconds and thirds) are easier (to sing) than larger intervals, and raises the question of whether they might also be easier to memorize.

Our main goal is to investigate the effects of familiarity and difficulty of melodies on our cognitive ability to reproduce and to retrograde such melodies at first hearing. It should be noted that we do not attempt to define the general notions of familiarity and difficulty in music, but instead we identified two particular aspects that seem to capture a fragment of these general notions. By adopting diatonic and chromatic scales as representatives of more or less familiar melodic contexts, we are constraining our experiment in well-defined ways, enabling us to question whether (this aspect of) familiarity influences melodic span measures. Also, by comparing the memorization of melodies made up of only small intervals to general melodies without interval constraints we may have a glimpse at the effect of melodic difficulty on our working memory.

Although the answer to these questions may appear self-evident for a practising musician, we intend to give objective, experimental answers to these questions. These answers, it should also be noted, are assumed to depend on history and culture, and ours are no exception, since we work within the biased boundaries of our experimental population. Our efforts are not directed to uncovering universal or innate facts about human cognition, and we make no claim to universal validity. Any such claim would have to be verified by crosscultural or transhistorical experimentation.

Another goal of this text is to discuss the differences and similarities between verbal and melodic memorization, and their possible implications for the structure of the working memory model. By comparing performance measures, in both forward and backward span tests, for sequences of digits and tones, it is possible to better understand the underlying mechanisms that comprise working memory. Specifically, we add a few experimental facts to the discussion of whether there might be a separate short-term memory component for dealing with tonal information [4, 5, 6, 7].

This paper is organized as follows. We introduce the cognitive model of working memory in section 2, and describe the methodology for constructing and applying the experiment on human subjects in section 3. We discuss the computational analysis of experimental data, the statistical analysis for hypothesis testing, and the experimental results in cognitive terms in section 4, and final remarks and pointers to future research are given in section 5.

2. THEORETICAL FRAMEWORK

The working memory model proposed by Baddeley and Hitch [2] in 1974 consists of three interconnected components (see figure 1), namely the *central executive*, the *phonological loop* and the *visuospatial sketchpad*. The system formed by these interconnected components is supposed to account for short-term storage and real-time processing of incoming information, and is vital for higher cognitive functions such as reasoning, planning and communication. A fourth component named *episodic buffer* was later added by Baddeley [3], but its discussion lies outside the scope of this paper.

According to this model, the phonological (also called articulatory) loop is responsible for short-term storage of auditory information and is capable of maintaining items in memory, for instance by using a subvocal rehearsal process. The visuospatial sketchpad (or scratchpad) allows for temporary storage and manipulation of visual and spatial information. The central executive is the attentional focus of the system and is responsible for controlling and coordinating the other subsystems, allowing for the recollection of recent experience and the symbolic manipulation of recollected items [3]. Both visuospatial and phonological subsystems are supposed to interact with long-term memory components, such as language and visual semantics, that may aggregate meaning to items in working memory.

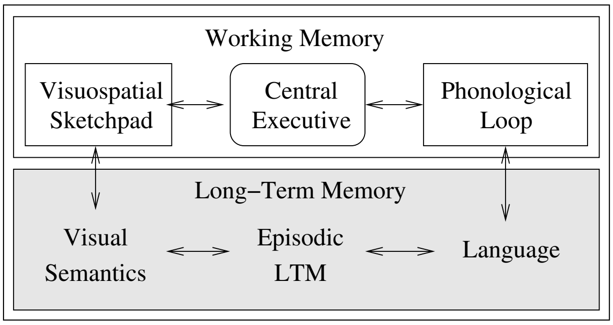

Figure 1. Baddeley and Hitch's Working Memory model.

A verifiable characteristic of these subsystems is the fact that they have limited storage capacity, which can be measured by *digit span tests* [8]. These tests consist of presenting random digit sequences of increasing length and asking for immediate reproduction. The forward digit span of an individual is defined as the maximum length of a sequence he or she is able to correctly reproduce; usually two sequences are tried for each length, to account for slips of attention or other disturbing factors unrelated to memory capacity. The backward digit span reflects an individual's ability to correctly reproduce a sequence of digits in inverse order, and measures the working memory capacity of simple symbolic manipulation of recently-presented items.

One interesting aspect of span measures is the fact that they are highly dependent on several aspects of the nature of information being presented. For instance, digit span measures differ significantly across languages, probably due to differences in word size, phonetic similarity and semantic context [9, 10]. The effect of these differences can be examined by measuring memory spans for sequences of words of controlled size and phonetic content, or for sequences in different languages using bilingual subjects. For these subjects, the measured span of recollection of sequences of numbers or words in their first language is higher than spans in their second language [10, 11, 9]. According to Thorn & Gathercole [9], the maintenance in the phonological loop of familiar sound patterns of a well-known language benefits from lexical and sublexical knowledge to complement mental representations, and is thus more effective for the first language than for the second language in bilingual subjects. These results suggest that the storage of items in the phonological loop is influenced by semantic and phonological long-term memory, as proposed by Ardilla [10].

Although phonetic and semantic aspects have been extensively studied within the working memory model for verbal items, purely tonal information have received considerably less attention in the literature [4, 5, 6, 7]. These works are concerned with *recognition tests*, where an isolated tone (stimulus) is presented and after a few seconds is compared to another tone (target). The insertion of irrelevant and unrelated tones between stimulus and target is known to degrade performance in tone recognition tests, whereas in digit recognition tests with irrelevant digits inserted between stimulus and target the decay in performance is barely noticeable [5]. This suggests that mechanisms of melodic and verbal storage might be independent.

One way of tackling this difficult question is by studying differences in memory span performance for melodic sequences both in direct and inverse order. One measure of performance degradation of backward spans with respect to direct span measures for the same information type is the *span index*, defined as the relative difference between forward and backward span measures. Different information processed by the same underlying mechanism would probably suffer from comparable degradation when passing from forward spans to backward spans, whereas significant differences in span index suggest that the underlying mechanisms might be different.

3. EXPERIMENTAL METHODOLOGY

In this section we present the methodology used in our experiments. The level of details offered should enable the realization of similar experiments with different populations and the comparison of both quantitative and qualitative results. We discuss the methodology in three stages. In section 3.1 we discuss the generation of the melodic sequences with varying levels of difficulty and familiarity (as discussed in section 1). In section 3.2 we discuss the prerequisites for individuals participating in the experiment, and also the first steps in selecting a reasonable population. The final application of the span tests is discussed in section 3.3.

3.1 Sequence Generation

The generation of data used in the melodic span tests is a crucial step in setting up the experiment, because the several sequences should reflect the relevant aspects of the questions we would like to answer. As discussed in section 1, we want to compare the difference in span performance in a more familiar and in a less familiar melodic context, as well as in a constrained, less difficult interval context and in an unconstrained, more difficult interval context. The defining attributes for these musical contexts in this particular experiment is as follows:

Familiarity: sequences are generated either within a single diatonic scale (e.g., C major) or within a chromatic scale.

Difficulty: subsequent tones in a sequence are generated either with constrained intervals (up to a major third upwards or downwards) or without any interval constraints.

These categories might be easily extended to consider other scales (e.g., pentatonic or quarter-tone scales) and other levels of difficulty (e.g., leaps up to a fifth or up to an octave), but the duration of the tests increase correspondently, and can easily become unbearable for the experimental subject. The average duration of the current experiment for each subject was about 90 minutes.

For each of the four combined contexts (more/less familiar and more/less difficult) a list of sequences of ascending length is generated, starting with 2 notes and going up to 10 notes, and always in pairs (2 sequences with N notes, for N=2,. . .,10). Since these tests require the subject to sing a melodic sequence, care should be taken with respect to the range of allowable tones. Each individual voice has its own tessitura, but in order to achieve uniformity of data and results some sort of compromise must be reached. We adopted a common range for female voices of [C4. . .C5], that correspond roughly to the intersection of soprano and alto registers (considering non-professional singers), and correspondingly the range of [C3. . .C4] for male voices. This corresponds to using up to 8 distinct tones in diatonic sequences and up to 13 distinct tones in chromatic sequences.

In order to be able to compare the effects of these contexts to what happens in similarly constrained verbal contexts, each melodic sequence was used to create a corresponding numerical sequence, by adopting the translations C4=1, D4=2, . . ., C5=8 for diatonic sequences and C4=1, C#4=2, . . ., C5=13 for chromatic sequences (and analogously for male voices). This way, we may also verify whether such restrictions on the number of symbols and internal structure of the sequences have some impact in span performance measures of numerical sequences.

Three additional constraints that appear in digit span tests were added in the sequence generation in order to avoid redundant sequences, which might be easier to recall due to the effect of *chunking* [3]:

- tones in a sequence can only reappear if strictly necessary (i.e., if sequence length > # of distinct tones used), and in such case there should be at least four distinct intermediate tones between repeated tones.
- sequences with large monotonic subsequences (e.g., 5 or more successive upward or downward steps) or with few direction changes (e.g., less than 2 breakpoints in a sequence with 7 or more tones) should be discarded.
- sequences with a large common subsequence ($N \geq 3$) with respect to the previous sequence in the same set (or a large repeating subsequence within it) should also be discarded.

Sequences to be used in inverse order were independently generated, instead of reusing direct order sequences, to avoid long-term memory collaboration. A total of 144 melodic sequences were thus generated, and the same amount of numerical sequences were obtained by direct translation.

Subsequently, all sequences were converted to audio, to guarantee that every individual is exposed to the same stimuli. Tones were synthesized as suggested in [7], by using plain sine waves, with 0.1 sec fade-in and fade-out ramps and a total duration of 0.5 sec per tone, followed by 0.5 sec of silence. Numbers were recorded using both female and male voices and sequenced in order to keep the same duration of 1 sec between the starts of consecutive numbers.

3.2 Population Requirements

A first requirement for any individual participating in this experiment was already stated in the previous section. Since responses are collected via singing, the individuals have to be able to sing; more precisely, we need to be sure that each individual participating in the experiment has the ability to hear a tone and reproduce it correctly, within a reasonably defined tolerance.

For our experiment we considered a population of volunteers that consisted of amateur choir singers and music undergraduate students. This may be viewed as a heterogeneous population, since they show significant differences in musical background, singing skills and even musical memory skills (since those without sight-reading skills usually rely only on their memory for acquiring repertoire). With all their diversity, they generally satisfy the two most important aspects in defining the population for this experiment, which are (1) the common exposure to western popular and classical music and (2) the ability to sing in tune.

We defined a tuning test to be applied before the actual span tests, which consisted of hearing tones and reproducing each one immediately after hearing it (no sequence memorization required). We used a 12-tone row (taken from Schoenberg's Variations Op. 31) for this test, and only individuals who reproduced the 12 tones correctly would be considered for the final experiment.

The tolerance used to decide whether a tone has been correctly reproduced is also a critical point of the experiment. It should be noted that the experiment tries to grasp something that lies inside the subject's mind (i.e., in his memory), but the empirical data is modulated by his/her vocal skills. In an attempt to overcome this difficulty, we accept as correct any tone within a quarter-tone distance from the target, even if just in passing (in the case of an unstable vocal emission). More details are given in section 4.1.

3.3 Span Tests

An experimental session consists of a short explanation about the nature of the experiment and the format of the tests, after which the volunteer reads and signs a written informed consent to become part of the experimental research. This is then followed by the tuning test, and after that the actual span tests. The application of the 144 melodic span tests (plus 144 numerical span tests [1]) as defined earlier is divided into categories of similar data in ascending length order, such as "random diatonic sequences with restricted intervals in direct order", and so on. The ordering of these categories needs to be balanced, by using several distinct permutations of the categories, in order to cancel out the effects of fatigue and progressive familiarization of experimental subjects with the tests.

As in the case of classic digit span tests, each category of sequences of ascending length has 2 distinct sequences for each length value, and the span of a subject for that particular category is defined as the largest length N for which the subject correctly reproduces at least one of the sequences for all lengths up to N. This flexibility is supposed to account for distraction, singing mistakes, and other disturbing factors not necessarily related to an individual's working memory system.

Forward span measures correspond to span values for sequences that were supposed to be reproduced in direct order (i.e., as heard). Analogously, *backward span measures* correspond to span values for sequences that were supposed to be mentally reversed by the subject before being sung back.

The presentation of stimuli is always made using headphones to minimize external interference, and all responses are recorded using a microphone. In our tests, stimuli were organized for individual presentation in a computer with an external M-Audio MobilePre USB soundboard, and all recordings were made with a Samson C15 Studio condenser microphone using 16 bit samples and 44.1 kHz sampling rate.

4. EXPERIMENTAL RESULTS

The experiment described in the previous section was conducted with 13 volunteers, 8 amateur choir singers and 5 music undergraduate students. Of these, 10 volunteers passed the tuning test and were used in the analysis. The other 3 volunteers were amateur choir singers who had borderline tuning results (exactly 1 tone off by a semitone) and were discarded. Such borderline results might be attributed to distraction or other factors unrelated to perceptual or singing skills, and may be futurely included in analysis as a separate population.

The recordings were automatically analyzed and semi-automatically graded, giving a span measure for each individual and for each category of sequences (as discussed in the previous section). These measures were then submitted to statistical Analysis of Variance (ANOVA), out of which our original hypotheses were put to test. These steps and some cognitive remarks are given in the subsequent sections.

4.1 Analyzing the Recordings

All recordings were analyzed using a monophonic transcription audio system specially tailored for this experiment. The application context departs from traditional automatic transcription problems in several aspects, such as irrelevance of precise rhythmic information, a priori knowledge of timbral structure (voice), and silence as a separator of relevant events, to name a few. This characterizes a relatively simpler transcription subproblem, which is solved

[1] We use the term *numerical span* instead of the usual *digit span* because in our sequences items may be composed of two digits, and this is likely to affect span measures causing them to differ from well-known digit span values.

by a four-step method described below.

The first step of the analysis was based on [12]. Recorded signals were divided in frames of 1024 samples and a peak-detection strategy was applied for each frame, creating a set of candidate spectral peaks. The accuracy of peak estimates was improved using signal derivatives [13], and F0 estimates for each frame were obtained by maximizing the cumulative harmonic energy of each candidate peak [12].

The second step of the analysis consisted in filtering out spurious results of the first step by median filtering F0 values and subsequently marking nearly-silent frames as event separators. This step produced a nearly stable F0 profile for each isolated utterance.

The third step consisted in transcribing these F0 profiles into symbols in a 24-step quarter-tone scale. This was done in order to correctly identify tones that were off by half a semitone, which should be considered correct (see section 3.2). This rounding-up to a 24-step scale involves a round-up error of the order of 1/2 of a quarter-tone, or 1/8 of a tone, and so the total tolerance adopted for this analysis was actually 3/8 of a tone, which is not so much if natural vibrato is taken into account.

The last step of the audio analysis consists of grouping up those symbols of the 24-step quarter-tone scale corresponding to a single profile and translate them into pairs (N,P) where N stands for a possible note (such as C#3 or B4) and P is the percentage of time of that profile for which the note could be accepted as N. For instance, an output like the following

```
----------------------------------------
Event Detected:          Intensity=411.783
Start=0.673s  End=1.196s  Duration=0.522s
Possible Notes: D 4 (99%), C#4 (29%)
----------------------------------------
Event Detected:          Intensity=515.923
Start=1.428s  End=1.974s  Duration=0.546s
Possible Notes: C 4 (99%), B 3 (60%)
----------------------------------------
```

states that the first note could be accepted as a D4 for 99% of that utterance's duration, but it could also be accepted as a C#4 for 29% of the time (it might be the case that the note was a little bit flat during attack or decay), whereas the second note could be either C4 or B3 (because F0 values were in between these two notes for 60% of the time). So the output of this analysis can be seen as a probabilistic transcription taking the tolerance of 3/8 of a tone into account.

All recordings were semi-automatically graded by this transcription system. By that we mean that conversion of recordings into span measures has been double-checked by a musically trained person. This was done for two main reasons: (1) to minimize the possibility of automatic transcription errors being transferred to the statistical analysis, with an impact in cognitive results; and (2) to gather extensive subjective evaluation about the transcription system, by applying it to over 800 recorded notes, and verifying that correct notes (according to a musically trained person) were always identified by the transcription system with $P>33\%$.

4.2 Statistical Comparison of Span Results

The output from the previous analysis is a set of numerical measurements for each individual and each test category. For simplicity, these categories were labeled with short names such as 7 and 12 for diatonic and chromatic span measures, and 3 and X for the categories related to difficulty (3 = intervals constrained to at most a major third, and X = no interval constraint). This data was submitted to a Repeated Measure Two-Way ANOVA on the effects of familiarity and difficulty, and post-hoc Newman-Keuls tests when necessary, using Statistica$^{\copyright}$ r5. Each possible comparison between groups of measures that might be statistically different has a corresponding significance level p, and small values of p (typically $p<0.05$) are interpreted as indicating a real difference between groups.

Figure 2 shows the averages and standard errors for the melodic span measures in direct order, or forward melodic spans (FMS), in all four categories.

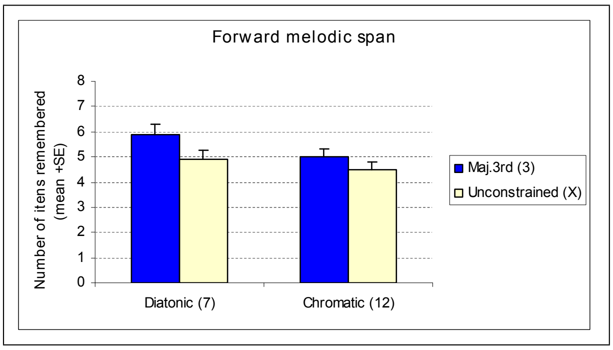

Figure 2. Forward melodic span measures.

The average forward melodic span for category 7_3 (5.9 notes) was significantly higher than the others (FMS(7_X)= 4.9, FMS(12_3)=5.0 and FMS(12_X)=4.5), with a significance level $p<0.023$. We may assume that the smaller number of items combined with a simpler internal structure does in fact ease the memorization task. Pairwise comparisons between the other 3 categories do not show statistically significant differences. This is not equivalent to saying that there are no differences, but simply that the experimental data for this population does not allow such conclusions to be drawn with reasonable confidence. A larger population might improve significance levels, allowing other hypotheses of pairwise comparisons, such as span(12_3)>span(12_X), to be confirmed or refuted.

With a two-way ANOVA we can study the effects of the familiarity (scale) irrespectively of difficulty (constrained or unconstrained melodic leaps), by combining all measures for the diatonic scale (7_3 and 7_X) and statistically comparing this group of measures to the results for the chromatic scale (12_3 and 12_X). This comparison allows us to conclude that average measures for the diatonic scale (FMS(7)=5.4) are significantly higher ($p<0.018$) than measures for the chromatic scale (FMS(12)=4.75). Comparing the two levels of difficulty irrespectively of familiarity leads to a similar conclusion, i.e., average measures for constrained sequences (FMS(3)=5.45) are significantly higher ($p<0.026$) than for unconstrained sequences

(FMS(X)=4.7).

It is interesting to compare these results to the corresponding span measures for numerical sequences that were built after melodic sequences by direct translation. Figure 3 shows the results of these tests. The labels (7), (12), (3) and (X) have been maintained, although in this context they only reflect the amount of allowed numbers (8 or 13) and the allowed differences between adjacent numbers.

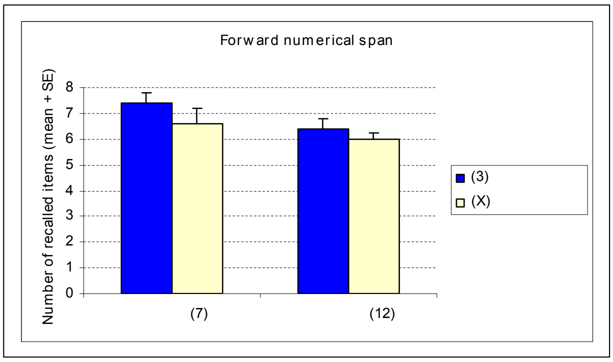

Figure 3. Forward numerical span measures.

Considering all four categories, forward numerical spans are higher than forward melodic spans ($p<0.020$), which means that sequences of numbers are more easily recalled than melodic sequences in the context of this experiment.

Here we also observe the same combined results as before, namely FNS(7) is significantly higher than FNS(12) ($p<0.006$), and FNS(3)is significantly higher than FNS(X) ($p<0.024$). This raises some important questions about the interpretation of melodic span results, which will be addressed in section 4.3.

We now turn to melodic spans in inverse order, or backward melodic spans (BMS). Figure 4 shows averages and standard errors for this experimental data.

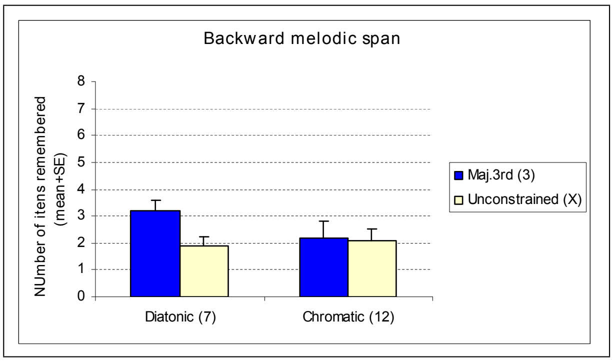

Figure 4. Backward melodic span measures.

The only worthy comparison here is between BMS(3)= 2.7 and BMS(X)=2 which refer to difficulty levels, but the significance level p=0.094 is higher than 0.05, which means that the confidence in this comparison is relatively low. This might be confirmed with a larger population.

It should be mentioned that these backward span measures are affected by the presence of several zeros corresponding to subjects who couldn't reproduce any sequences in reverse order (sequences start with 2 distinct tones). This, combined with many other low results (BMS=2) contributes to what is called *floor effect*, which has important consequences for statistical analysis of these data.

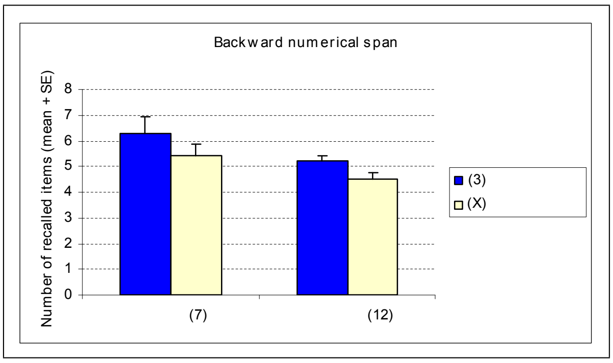

Figure 5. Backward numerical span measures.

Figure 5 shows backward numerical span measures. This data allows the conclusions that BNS(7)=5.85 is higher than BNS(12)=4.85 ($p<0.017$), and that BNS(3)=5.75 is higher than BNS(X)=4.95 ($p<0.019$), or in other words, both the smaller number of symbols and the simplified interval structure of the sequences do in fact help the memorization and mental reversal of sequences of numbers.

It can also be drawn from this data that backward numerical span measures are higher than the corresponding backward melodic spans ($p<0.005$). This means that retrograding melodic sequences is in fact much more difficult than reversing numerical sequences, and the confidence level of this conclusion is high.

The next section focuses on possible cognitive interpretation of the above quantitative and qualitative conclusions.

4.3 Cognitive Aspects

We shall first address the differences and similarities in forward span measures for melodic and numerical sequences. We concluded in section 4.2 that numerical span measures were generally higher than melodic span measures. This could be explained by the many associations that numbers in working memory have with long-term memory knowledge, such as visual and linguistic alternative representations. A similar phenomenon has been observed in individuals with absolute pitch, who resorted to verbal strategies to achieve a higher melodic span [14].

Another interesting comparison is the fact that the restricted contexts (7 and 3) did increase forward span measures with respect to less restricted contexts (12 and X), both with melodic and numerical sequences. This raises the possibility of a single explanation accounting for both phenomena, which might not be an exclusively musical explanation. Items (numbers, pitches) that are close to one another in their respective representation spaces might be more effectively combined into larger chunks (subsequences, motifs) in the working memory, effectively allowing a larger number of items to be stored.

It has been observed that the number of symbols (8 or 13) also affect span measures. This effect might be intuitive in the numerical domain, since some numbers are represented by two digits and might also have a comparatively larger mental representation. But in the musical domain we have been looking at those categories (7 and 12) as representatives of more or less familiar contexts. It might be argued that a single explanation (number of symbols) would account for both observations. We would counterargue that chromatic sequences with length less than 8 also have less than 8 distinct symbols, so non-diatonic 8-element subsets of a 13-element chromatic scale already appeared in our experiment; the only difference is the fact that these 8-element subsets are not fixed within each category. An experiment might be made using other 8-element fixed subsets of a 13-element chromatic scale to provide a more well-founded comparison.

In backward melodic span measures we observed a floor effect that make it more difficult to draw qualitative conclusions from statistical analysis. It might be the case that chunking of close elements within a musical scale make the process of reversal of a sequence easier. In any case, by comparison with the reversal of numerical sequences, musical retrogradation of unheard melodic sequences appears to be a very difficult task.

It is interesting to notice that backward digit span measures are affected by restricted contexts such as 7 (8 instead of 12 symbols) or 3 (small rather than large intervals). This suggests that chunking of information in working memory is probably more effective in reversing numerical sequences rather than melodic sequences.

It might be wondered about the effect which training would have in both tests. Reversal of numerical sequences does not appear to be a frequently applied task in elementary school, and the same could be said about melodic retrogradation without the aid of a writing pad. Yet the results suggest that dealing with numbers in working memory is naturally easier than dealing with notes, in the sense that our population was not specifically trained for neither of these tasks.

These differences in behavior of backward span measures with respect to forward span measures are made more clear when they are expressed by relative differences or *span indices*, defined as

$$\frac{\textit{(forward span - backward span)}}{\textit{(forward span)}}.$$

Figures 6 and 7 show these values for melodic indices and numerical indices, respectively.

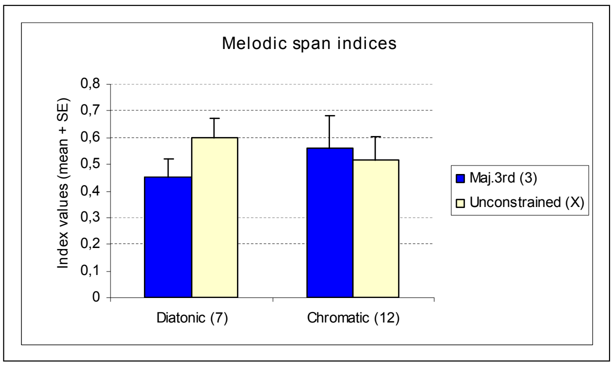

Figure 6. Indices for melodic span measures.

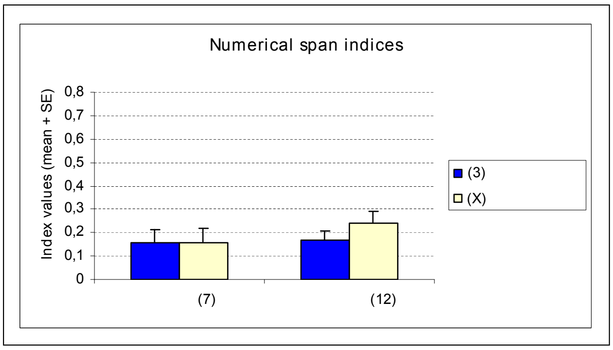

Figure 7. Indices for numerical span measures.

These values reflect the relative difficulty in mentally reversing sequences with respect to simply reproducing them in direct order. Statistical analysis allow the conclusion that numerical indices are higher than melodic indices with a significance level $p<0.0006$.

It is interesting to compare these indices to numerical indices of other languages. For instance, digit span indices for English, Spanish, Hebrew and German are in the range [0.09,. . .,0.26] [10, 15, 16], and this range also includes all four numerical span indices that we obtained.

On the other hand, digit span indices for Mandarin are relatively higher, around 0.48±0.05 according to Hsieh & Tori [17]. This value is closer to what we obtained as melodic span indices. It might be argued that Mandarin is a tonal language, meaning that pitch variation within a phoneme is a component of semantic value, and so even the task of remembering numbers (or reversing them) requires some attention to melodic profile.

These differences suggest that the underlying mechanisms for verbal and tonal processing might be different, as suggested by other authors [7]. Baddeley's working memory model includes separate components for visuospatial and phonological information, but does not distinguish between phonological information with verbal content or purely acoustic information. By observing the differences in numerical and melodic span indices we could consider a subdivision of the phonological loop into two components responsible for verbal and acoustical material, or even the existence of a component for acoustic processing which is separate from the phonological loop.

5. CONCLUSION AND FURTHER RESEARCH

This paper has brought experimental facts about human music cognition, which might be relevant for computational psychoacoustics and for the development of cognitive models for automatic feature extraction. We have studied the impact of familiarity and difficulty in the task of memorizing melodic sequences, by adding simple constraints to the generation of test sequences.

We observed that both familiarity and difficulty (in the sense defined in section 3.1) contribute to higher forward melodic span measure. A similar finding in forward numerical span measures adds to the understanding of the melodic results in two ways: it provides a possible explanation to measure differences related to difficulty as a consequence of chunking, and it also raises the question of whether the number of symbols alone would be responsible for the observed differences with respect to what we called familiarity.

Observing behavioral differences in backward numerical and melodic span measures, and specially comparing span indices to other well-known experiments, we suggest that the underlying mechanisms for dealing with verbal and acoustic information in working memory are probably not the same, since a similar mechanism operating similarly on both information would not display the observed levels of degradation in backward spans with respect to forward span measures.

The experiment described here can be easily extended and applied to other population groups. Some of the factors that may contribute to relevant findings are: the size of the population, considering other groups such as professional singers or non-singer professional musicians, and also considering other levels of familiarity or difficulty or even other aspects of melodic sequences not contemplated here.

Future work may also combine this type of experiment to neuroimaging techniques to help mapping cognitive subsystems of the working memory model to particular activation areas in the human brain. Some studies that follow this idea are the localization of regions involved in recognition tests with melodic material using PET scans [7], and the localization of areas involved in the subvocal rehearsal strategy of the phonological loop for verbal and melodic material using fMRI [18].

ACKNOWLEDGEMENTS

The authors would like to thank FAPESP, CNPq and AFIP.

6. REFERENCES

[1] Atkinson and Shiffrin, "Human memory: a proposed system and ist control processes," in *K. W. Spence and J. T. Spence (Eds), The Psychology of Learning and Motivation*, Academic Press, 1968.

[2] A. D. Baddeley and G. Hitch, "Working memory," in *The Psychology of Learning and Motivation*, vol. 8, pp. 47–90, 1974.

[3] A. D. Baddeley, "Working memory and language: an overview," in *Journal of Communication Disorders*, vol. 36, pp. 189–208, 2003.

[4] S. Berti, S. Münzer, E. Schröger, and T. Pechmann, "Different interference effects in musicians and a control group," in *Experimental Psychology*, vol. 53(2), pp. 111–116, 2006.

[5] D. Deutsch, "Tones and numbers: Specificity of interference in immediate memory," in *Science*, vol. 168, pp. 1604–1605, 1970.

[6] R. H. Logie and J. Edworthy, "Shared mechanisms in the processing of verbal and musical material," in *Imagery II, edited by D. G. Russell, D. Marks and J. Richardson*, pp. 33–37, 1986.

[7] R. J. Zatorre, A. C. Evans, and E. Meyer, "Neural mechanisms underlying melodic perception and memory for pitch," in *The Journal of Neuroscience*, vol. 14(4), pp. 1908–1919, 1994.

[8] J. T. E. Richardson, "Functional relationship between forward and backward digit repetition and a non-verbal analogue," in *Cortex*, vol. 13, pp. 317–320, 1977.

[9] A. S. C. Thorn and S. E. Gathercole, "Language-specific knowledge and short-term memory in bilingual and non-bilingual children," in *The Quarterly Journal of Experimental Psychology*, vol. 52A(2), pp. 303–324, 1999.

[10] A. Ardilla, "Language representation and working memory with bilinguals," in *Journal of Communication Disorders*, vol. 36, pp. 233–240, 2003.

[11] A. S. C. Thorn, S. E. Gathercole, and C. R. Frankish, "Language familiarity effects in short-term memory: the role of output delay and long-term knowledge," in *The Quarterly Journal of Experimental Psychology*, vol. 55A(4), pp. 1363–1383, 2002.

[12] A. Mitre, M. Queiroz, and R. Faria, "Accurate and efficient fundamental frequency determination from precise partial estimates," in *Proceedings of the 4th AES Brazil Conference*, pp. 113–118, 2006.

[13] M. Desainte-Catherine and S. Marchand, "High precision fourier analysis of sounds using signal derivatives," in *Journal of the Audio Engineering Society*, vol. 48(8), pp. 654–667, 2000.

[14] M. E. Benassi-Werke, "Memória operacional para tons, palavras e pseudopalavras em músicos," in *Anais do SIMCAM4 - IV Simpósio de Cognição e Artes Musicais*, pp. 1–9, 2008.

[15] H. Silver, P. Feldman, W. Bilker, and R. C. Gur, "Working memory deficit as a core neuropsychological dysfunction in schizophrenia," in *American Journal of Psychiatry*, vol. 160(10), pp. 1809–1816, 2003.

[16] T. Merten, P. Green, M. Henry, N. Blaskewitz, and R. Brockhaus, "Analog validation of german-language symptom validity tests and the influence of coaching," in *Archives of Clinical Neuropsychology*, vol. 20, pp. 719–726, 2005.

[17] S.-L. J. Hsieh and C. D. Tori, "Normative data on crosscultural neuropsychological tests, obtained from mandarin-speaking adults across the life span," in *Archives of Clinical Neuropsychology (in press)*, 2007.

[18] S. Koelsh, K. Schulze, D. Sammler, T. Fritz, K. Müller, and O. Gruber, "Functional architecture of verbal and tonal working memory: an fmri study," in *Human Brain Mapping*, vol. 30, pp. 859–873, 2009.

[19] W. L. Berz, "Working memory in music - a theoretical model," in *Music Perception*, vol. 12(3), pp. 353–364, 1995.