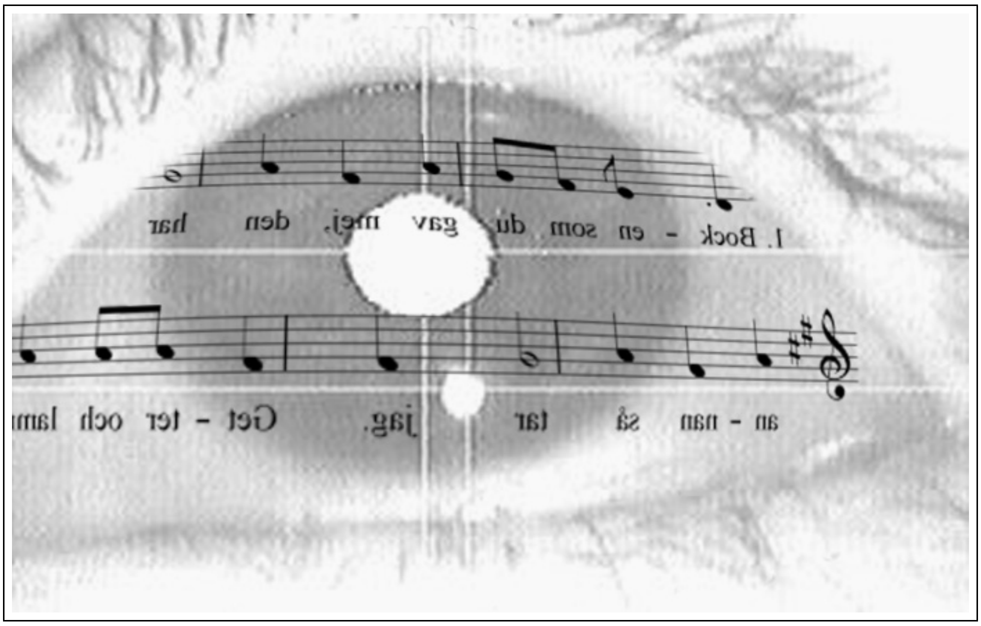
Eye movement in prima vista singing and vocal text reading


1. Bock - en som du gav mej, den har
an - nan så tar jag. Get - ter och lam

Per Berséus[1]
Lund University Cognitive Science

Undergraduate paper at D-level
April 17, 2002

[1] Supervisor: Kenneth Holmqvist, LUCS

Abstract

An eye-tracking device was used in this pilot study to measure eye movement during prima vista singing and reading aloud. The common term eye-voice span is questioned, and a new terminology is proposed for the description of relations between attention, point of fixation and vocal performance. The temporal distance between the point of fixation and the vocal performance was measured and proved to be larger in vocal language reading than in musical sightreading. This distance is also related to the eye-voice span and the span of the window paradigm. Regressive saccades were found to be less frequent than expected from previous research on silent language reading and music sightreading. It is suggested that the difference depends on obscure definitions of the term regression. Furthermore, a sheet displaying both text and notes was used to study the distribution of attention between text and notes, showing that attention was almost equally shared.

KEYWORDS: eye movement, eye-voice temporal distance, eye-voice span, music notation, prima vista, reading aloud, regressions, saccades, sightreading, singing

Contents

1 Introduction

1.1 Inspiration

When eyes meet, a certain connection is created between two people. The eyes tell you something about a persons feelings and thoughts, and nothing in a face attracts attention as much as the eyes.

The reason why people look each other in the eyes is unfortunately nothing romantic and exiting about creating magical soul bonds, but something as boring as efficiency of information collection. The eyes can tell you about the cognitive processes of a person, they are a window of the mind.

Since vision is one of our main sources of information, people tend to use it efficiently, by looking at places where the maximum amount of information can be gathered (Yarbus, 1967). Thus, if you know what someone is looking at, you can say something about which information is important for the person at that moment.

Psychologists have developed eye-tracking devices, which have been used to explore the eye movements during reading. For example, they can tell you about differences between the behaviours of fast readers and slow readers (Hyönä et al. in press). It is tempting for a slow reader to try to benefit by this research and mimic the eye movements of a faster reader, in an attempt to increase reading speed. However, there is no guarantee that the eye movement patterns of fast readers are the cause of their good speed. They might just as well merely be the result of other processes, in which case it would be a waste of time to try to change one's own natural patterns.

The same sort of false expectations could concern studies such as this one, where the eye movements of sightreading singers are subject to examination. Musicians around the world would surely love to learn about which visual patterns seems to be the most efficient, but currently the research of music reading cannot give anyone that sort of advise.

Instead, what eye-tracking measurements can result in, which also is the aim of this and numerous other studies, is knowledge about the way we perceive the world, the way we process this information and, hopefully, something about the way we think and feel. Then the eyes can truly become a window of the mind.

1.2 Motivation

This pilot study was designed to further explore the concept of reading. A lot of effort has been put into silent language reading research (there is an extensive review in Rayner, 1978), and a few studies have been performed on music reading (reviewed in Sloboda, 1984 and Goolsby, 1994a). Music reading is a promising field of research, and since there exists many similarities with language reading, comparisons may prove beneficial. However, no attempt has been made to comprise both music and language reading in the same study, an approach which appears superior when comparisons are to be made between these two modes of reading.

Thus, the study design is built around three types of stimuli, one purely textual, one purely musical and one combined stimulus of music with lyrics.

Since the choir singers who participated as subjects were to vocalize the musical stimuli, it seemed appropriate to have them vocalizing the text as well. However, the previous research of reading aloud is surprisingly sparse, and it is unclear to what extent silent reading results may be assumed to be valid for vocal reading as well.

An eye-tracking devise was used to record the eye movement, and different data treatment techniques were tried out. Due to the explorative nature of this pilot study, ecological validity was deemed more important than rigorous experimental control. Thus the quantitative results, and to a certain degree also the qualitative ones, must be verified by further research.

It would be of great benefit to interest both linguists and musicologists in this comparative research. In order to facilitate for different kinds of readers of this paper, no previous knowledge about eye movement or musical notation is presupposed. The basics of eye movement are summarized below. An introduction to standard Western musical notation is found in appendix A, and a review of research in language and music perception research in appendix B.

1.3 Vision and reading research

A review of the research conducted on eye movement and attention is a prerequisite for a fruitful discussion of the results in this pilot study. The focus is on the reading of music and language.

1.3.1 Eye movement

The qualitative research of eye movement stretches back more than a century, and its main characteristics are well documented (Goolsby, 1994a). The movements in reading and picture viewing constitutes mainly of *saccades*, swift ballistic movements during which the eye is blind. Visual information is acquired during the intermediate *fixations*, where the eye movement is minimal (Yarbus, 1967).

The distance covered by a saccade is measured by visual angle, where in reading the average distance amounts to approximately 2°, equalling a saccade duration of 25–30 ms (Rayner, 1978).

The saccades are not of random length or direction. Yarbus (1967) concludes that elements in a picture that attract attention are those who contain useful information. In langauge reading (i.e. text reading), few fixations occur at common words of little new information, e.g. the article “the” (O’Reagan, quoted by Sloboda, 1985), while in music reading larger saccades are made in areas where the duration of notes are predominantly shorter (Kinsler and Carpenter, 1995).

In text reading, the average fixation duration of a skilled reader is 200–250 ms (Rayner, 1978). In music sightreading the fixation duration is longer, around 400 ms (Goolsby, 1994a), as well as in picture viewing, where it is around 330 ms (Henderson & Hollingworth, 1999).

1.3.2 Eye-tracking

The eye movements can be monitored with an eye-tracking device, which keeps record over where the eyes are directed during the recording. This information tells you where on the stimulus the fixations are, as well as their duration. The sequence of fixations and intermediate saccades form a *scan path* (figure 1), which gives a good overview of eye movement behaviour.

1.3.3 The visual field

The visual cells are not homogeneously distributed over the retina, and only the light falling upon the *fovea* renders the really sharp image required for reading text or music. When a point of visual stimulus is said to be focused, the fovea of the eye is directed at an area of approximately 1° of visual angle surrounding that point.

This entails a physical limitation of eye-tracking measurements, since a subject during calibration may attend to one spot within the area of focus (the size of which is equivalent to a thumb-nail viewed from the distance of an arm’s length), and during a later fixation attend to another spot.

Although the image is not sharp outside this central area of the visual field, some information is apprehended from the periphery. In vision research, the visual field is divided into three areas, where the *foveal region* is within the central 2°, the *parafoveal region* encompasses 10° around the point of fixation, and the *peripheral region* is everything outside the parafoveal region (Rayner, 1978).

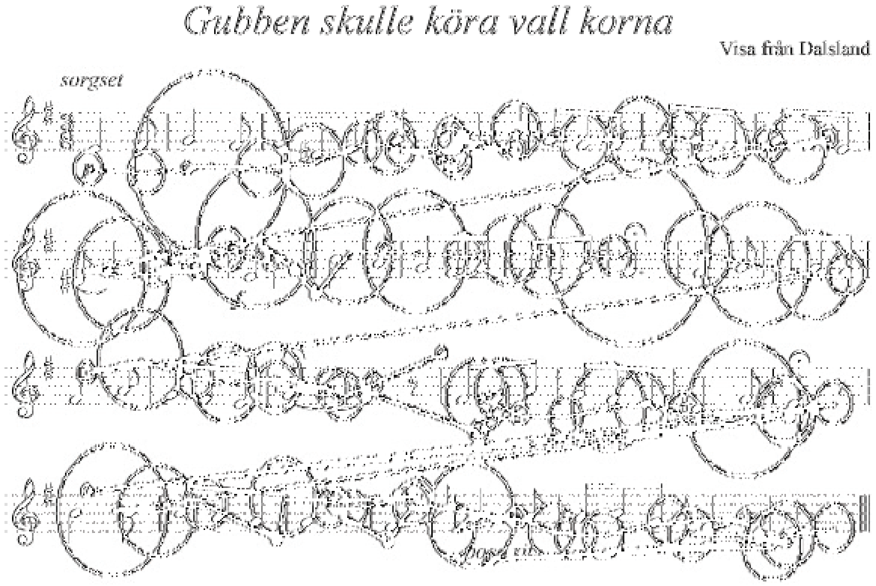


Figure 1: *This* scan path *displays fixations as circles with radii proportional to time. The saccades are straight lines. The scanpath of this figure was fitted to the music with a separate image manipulator program, hence the distorted circles and unprecise path.*

Peripheral vision is believed to have an important role in guiding the saccades, and the information collected during subsequent fixations may be integrated, thus facilitating reading and increasing the speed (Rayner, 1978). Kinsler and Carpenter (1995) remarks that the last fixation on a staff is halfway through the last bar, which is a clear manifestation that information collection occurs outside the foveal area. This behavior is also well know from language reading (Rayner, 1978). Experiments has been conducted by e.g. Rayner and MacConkie (ibid.) where ordinary text has been modified in different ways, except for in a *window*, a small area around the point of fixation. This window follows the point of fixation, and the idea is to observe how reading performance is affected when the window is small. According to Rayner (ibid.), the window paradigm experiments has shown that information about word length affects the reading further out in the periphery than word-shape or specific-letter information does. The details of words and letters appeared to influence the reading not more than 10-11 character positions to the right of the point of fixation, while information about word-length was not acquired more than 15 character positions from fixation.

1.3.4 Attention

Generally, the visual attention and the point of fixation follow each other closely. Occasionally, we look at something without attending it, e.g. one may read through a whole page without catching a single word. We can also attend to something without looking at it, which is when something is viewed in the corner of the eye, by peripheral vision (Hansen, 1994).

The question of how close the linkage between visual attention and eye movements really is remains to be settled. The evidence at hand suggest that a saccade always is preceded by a shift of attention to the target location and that the attention between saccades is divided between foveal and peripheral vision (Hoffman, 1998). Deubel and colleagues have in several experiments shown that the attention shift to the target location of the saccade may occur very early in the fixation, as much as 250 ms before the saccade is executed (Deubel et al., 1999).

Attention is not always focused to the same degree, which is the most apparent in picture viewing. This feature can be described with *the spotlight metaphor* (Holsanova, 2001). Like a spotlight, the gaze moves across a picture, and it sometimes zooms out to get an overview, sometimes zooms in to study certain details. As a matter of course, this effect must be much more limited in text reading, where the detail level of the letters is by far the most important, but may have stronger significance in music reading, especially after several encounters with the same musical material, when an overview could suffice.

Even though we do not always attend what we look at, the cognitive load in prima vista singing is high enough to avoid subjects looking right through the stimuli. The demands on temporal fluency in vocal reading is assumed to induce such a work load as well. It is assumed, in this study, that eye-movement recordings actually can say something about visual attention.

1.3.5 Perceptual span

While reading text or music, the eyes are usually ahead of the voice, or — said in a more straightforward manner — you need to see a word or a note before you are able to vocalize it. In reading research, this distance has been referred to as the *eye-voice span*. According to Sloboda (1974), the term was suggested by Levin and associates in 1970.

In eye-voice span studies, the illumination of the text read aloud by a subject is turned off when a pre-determined point of performance is reached (Rayner, 1978). The subject is asked to report the words seen but not yet pronounced at the time the light went out. This is taken as the eye-voice span.

Sloboda (1974) is careful to make a distinction between the eye-voice span and the *span of apprehension*, which denotes the total amount of visual information available at any time. The eye-voice span rather represents the amount of information which can be extracted from the span of apprehension before the visual trace decays, i.e. within a couple of seconds. The rate at which the information is extracted from memory is thus limiting the eye-voice span.

For normal prose readers, the eye-voice span is five to six words, and for piano players, the eye-hand span is five to six notes. The span may be extended if a phrase ending is a little further than this, and it is likely to contract when less than five words or notes remains in the phrase. Also, the eye-voice span decreases when the stimulus becomes more complex (Sloboda, 1984; Rayner, 1978).

Goolsby (1994b) uses the term *perceptual span*, which is more loosely defined as the region of a visual stimulus that can be seen during a single fixation. The term is also used by Rayner (1978) in a general sense. Rayner & Morris (quoted by Goolsby, 1994b) have remarked that, according to numerous studies, the perceptual span in silent language reading extends about 15 characters to the right of the point of fixation. Goolsby (ibid.) suggests that a temporal measurement of the perceptual span may perhaps be more appropriate, and he also compares his own result that the eye was 2000 ms ahead of the point of performance with an indication from Sloboda that musicians read 2 seconds ahead of performance. However, Goolsby used an eye-tracker to trace the eye-movement, and Sloboda's studies were based upon stimulus removal. These methods may not produce comparable results, and this confusion will be further addressed in section 2.2 below.

1.3.6 Eye movement in language and music reading

In ordinary reading, the eyes move essentially linearly from left to right, with saccades of different length. The visual process itself, where information is collected only during fixations, implies that some sort of visual buffer is needed in order to process data smoothly without interruptions. Evidence suggests that in such a buffer visual information is integrated between saccades (Rayner, 1978).

The saccades are usually directed from left to right, except for the *return-sweep* at the end of each line. Sometimes, though, this progressive sequence is interrupted by *regressions*, where a fixation occurs at a point in the text already traversed. The share of regressions in silent reading is approximately 10–15 per cent (Reichle et al., 2000). As for the reasons for regressions, Carpenter and Just (quoted by Rayner, 1978) reports that regressive fixations are sometimes made to the referent of a pronoun.

Goolsby (1994a, 1994b) has made eye-tracking recordings of prima vista singers and, essentially, the eye movement in music reading resembles that of language reading. His results show that the share of regressions was somewhere around 30 per cent, which is considerably more than in language reading. Goolsby (1994b) proposes, on the basis of saccade lengths, that good readers look ahead in the notes in order to get a preview, and use regressions to get back to the point of performance when needed, whereas poor readers have to use regressions to search for information amongst the notes already performed and then get back to the point of performance with a progressive saccade.

In music reading, the vertical component is of greater significance than in language reading (Sloboda, 1984). This is especially apparent in scores that consists of more than one staff, e.g the double-staffed piano music. An early study by Weaver (reviewed by Sloboda, 1985) has shown that pianists use different strategies to deal with the situation of not being able to fixate both staffs concurrently. In *chordal* music, where the relation between notes played simultaneously is important, the saccades displayed a zigzag pattern between the staves. In *contrapuntal* music, where the melodic sequences are of greater importance, horizontal lines of a few saccades were followed by return to the other staff. Sloboda (ibid.) remarks that the strategy seems to be to identify significant structures in successive fixations.

Goolsby (1994a, 1994b) noticed, in his study of prima vista singing, features of eye movement that seems to display the opposite patterns than those of language reading. In music reading, subjects tends to fixate on white space between notes and on bar lines of connected quavers, whereas in language reading few fixations occur between words (Rayner, 1978). Goolsby (1994b) suggests that the intervals are more important than the actual pitches of the notes, and therefore it is desirable to capture the relation between two notes in a single fixation. The fixations on bar lines connecting quavers may be valuable to facilitate grouping of notes in memory. Furthermore, Goolsby (ibid.) has shown that good readers in music have more but shorter fixations, whereas the opposite trend is found in language reading.

Good sightreaders have a larger eye-voice span than poor readers, a preponderance that decreases when the music confines less to traditional rules of tonality and harmony. The good readers seem to be able to take some advantage of their musical knowledge in the encoding of the musical information, making it more efficient (Sloboda, 1984). Recognition of common rhythmical and melodic patterns may constitute such knowledge, making it easier to group the notes in the encoding.

1.3.7 Preview and expectations

Because of the integrative nature of the visual buffer, readers get a *preview* of words and notes before they are fixated, through peripheral and parafoveal vision. Evidence reviewed by Hoffman (1998) suggest that this is the reason that the saccade size is adjusted in a such manner that words carrying little information, such as "the", are skipped.

Besson and Friederici (1998) point out that language and music share the common tendency to evoke strong *expectations*. These can be built on patterns and conventions at different structural levels. The fact that one often can guess which word would be the next one in a discontinued sentence might question whether we really make use of the peripheral vision after all: is our saccade guidance merely a matter of pure guessing? Sloboda (1974) discusses the issue and concludes that since incorrect performances often have much similarities with the notated music, the pure guessing is out of the question. Instead he advocates a theory of *sophisticated guessing*, where expectations can be used to reduce the number of possible words or notes to appear as the next one, thus facilitating perception and achieving higher speed.

Experiments were conducted by Shaffer (reviewed by Sloboda, 1985) on copy typists with texts appearing gradually as they copied the words. It showed that a preview of 8 letters was needed to maintain good speed in the copying task. This, and the already mentioned window technique studies, show that the preview of material further ahead is an important factor, alongside with expectations, for a successful performance.

1.3.8 Cognitive load

The concept of cognitive load is complex and hard to define. It seems to be of value, at any rate, to make one certain distinction, viz. that between how much information processing is required and the load on the cognitive system, i.e. how laborious the task is to a person. Many tasks are difficult and complex, but still does not exert any pressure on persons that are used to the situation.

In some studies (e.g. Pelz et al., 2000), fixation duration has been shown to be shorter during complex tasks, and it is often taken as a measure of cognitive load. Rayner (1978) quotes Wanat, who has shown that at places where the eye-voice span is long, there are fewer and briefer fixations, which might imply that the eye-voice span has to do with the cognitive load as well, or possibly the amount of information to be processed.

However, Recarte and Nunes (2000), explain that although the fixation times are shorter during high processing rates and when information is distributed over a wider area in a complex scene, it is difficult to draw valid conclusions, since it also is well known that fixation time increases when much information is to be collected from a target.

Generally, musical sightreading is a task of high cognitive load. Few choir singers are really comfortable with prima vista singing, and it demands a high degree of concentration. A circumstance that adds to the cognitive load is that musical reading seldom is prima vista reading. The music notation of a piece is read through many times during rehearsals, and a singer does not depend as much on the notation as a reader, who usually read a text only once.

1.3.9 Silent versus vocal text reading

Silent reading is a fairly recent invention. Until the tenth century, the natural reading situation was one reader attended by a group of listeners[1]. Reading included

[1] Alberto Manguel shares his contemplations on the subject in *A history of reading*.

listening, and when people read to themselves, they pronounced the words and listened to their own voice. Eventually, people realized that it was not necessary to vocalize the words, which greatly improved reading speed.

Indeed, silent reading is a more efficient mode. Today, the speed of silent readers amounts to somewhere around 300 words per minute (Reichle et al., 2000), and any serious attempt to keep up such a pace in vocal reading would end up in ridicule.

Little research has been made on vocal text reading. In a summary of eye movement in reading, Rayner (1978) mentions eye-voice span recordings by Wanat, where half of the sentences were read aloud and half silently. Reading aloud resulted in longer fixations, but the overall pattern was said to be essentially the same.

However, Hyönä and colleagues (Hyönä et al., in press) have made distinctions between four different reading strategies in silent reading, depending on the reading speed and the systematicness of look-backs to previous sentences. These are not likely to be applicable to vocal reading as well. This example illustrates that there clearly must be more to explore within the topic of reading aloud than can be inferred from silent reading data.

Unlike silent reading, vocal reading shares some common features with music performance. The most obvious one is the constraints on the temporal sequence of the performance. When music or text is articulated, the words and notes cannot be produced in any order, and the performance should be fluent, without unmotivated pauses.

1.3.10 Silent versus vocal music reading

Kinsler and Carpenter (1995) indicates that silent music reading is a neglected area. However, there are difficulties with such studies. It is hard to find some equivalent to the comprehension tests that may be used in silent reading studies, which makes it a complicated task to know whether the whole piece really is read through by the subject.

In addition, music reading is, as text reading once was, never implemented in silent mode by most readers. In silent text reading, prosody and phonology can be ignored in favor of content, but it is not evident that the core of musical meaning may be stripped in a similar fashion.

1.3.11 Neural representation

Do musical perception and linguistic processing share the common neural resources in the brain? The lack of interference between tones and words presented by Deutsch (1970) can be taken as evidence for two separate memory systems. Recognition of pitch was in her study strongly affected by six other tones incorporated in a five second retention interval. When the tones were replaced by six spoken numbers and a task to later recall those numbers, the decrement of pitch recognition was minimal. Also, the pitch recognition task had no decrement on the number recalling task. The immediate memory for pitch must be subject to a large but highly specific interference effect, and Deutsch (ibid.) furtermore has reason to conclude that it is unlikely that we remember musical sequences by storing the absolute pitches, but instead store music in a recoded form.

However, Patel (1998) has by means of event-related brain potential (ERP) been able to show that the brain produces similar waveforms when a subject is confronted with musical and linguistic incongruencies. To consolidate this result with seemingly opposing findings he suggests a view where the syntactic processing of music and language rely on different cognitive operations but on the same neural resources. Different cognitive operations are used in language and music, but the structural linking between the current stimulus element and the previous ones uses the same neural resources. He concludes that music and language thus are different windows of the syntactic capabilities of the human mind.

Also, Patel and Peretz (1997) reviews results revealing that, in long term memory, music and text appear inseparable. It is difficult to remember the melody of a song without remembering the words, and vice versa. They advocate a view where music and language are not seen as independent mental faculties, but labels for complex sets of processes, some shared, some different. For example, neurophysiological evidence suggests (ibid.) that pitch employs the same resources, but tonality is specific to music.

1.4 Tonality

Tonality is the common situation where the relations between notes put one certain pitch in focus. In traditional music, which pitches belonging to a certain scale is determined by the musical mode, where the distinction between major and minor scales is central. Besson and Friederici (1988) reviews evidence from Pechmann that pitch is not as important as mode in a retention task.

The inclination to stay within a scale, is beautifully illustrated by the experiments conducted by Dowling (1978). The results of these show that *transpositions*, where every note in a melody was repeated with a displacement of the same number of half tones, were not separable from *tonal answers*, where a melody line was repeated with a displacement of the same number of note steps within the key.

Sloboda (ibid.) concludes on the basis of this that a melody coding is probable, where musicians in their memories store the number of scale steps rather than precise pitch distance. This also implies that important information can be inferred merely from the *contour* of a melody, i.e. the visually evident sequence of the vertical placement of notes, which should be fairly easy to catch with peripheral vision.

1.5 Notation

Rayner (1978) issues a warning that conclusions from picture viewing data does not necessarily account for processes in language reading, e.g. he adduces two major reasons not to overemphasize the significance of peripheral and parafoveal vision in the guiding of the saccades. Firstly, picture viewing is more explorative than text reading, and the informative parts of a picture reveal themselves more easily in pictures than in text. Secondly, the visual information in pictures is of a more directly visual type than that in text. The text is based on convention and its content is of syntactic and semantic type. The mapping between a picture of a tree and the tree concept is more fundamentally direct than the mapping between the word “tree” and the concept it stands for. Besson and Friederici (1998) points out that words carry meaning by convention, while music is more self-referential, i.e. tones having meaning mainly by reference to the preceding ones. A music notation stimulus is more direct than a text stimulus mainly in the pitch dimension: higher frequency is notated higher up in the staff. This suggests that peripheral vision may be of greater importance for pitch than for other musical features.

Sloboda’s (1985) hypothesis about only the number of scale steps being stored in the pitch memory is suggestively reflected in music notation, since the close mapping between vertical placement of musical notes and their pitch deteriorates and becomes more conventionalized for notes which do not belong to the key but are placed on the same line or space in the staff.

Sloboda (ibid.) remarks that grammatical structure is displayed more clearly in notation than in performance. Words and notes are separated in notation, but not so in performance.

The main temporal flow in language and music is represented similarly by progression to the right (in Western music and language notation, i.e.). However, as Kinsler & Carpenter (1995) points out, the tempo variations are in vocal language reading arbitrary, and pauses may be inserted more or less at will, while the demands in music reading are stricter. They suggest that the more strict temporal directions in music notation ought to submit the eye movement to stricter cognitive control than is displayed by the somewhat irregular saccades in text reading.

Why are there greater demands of strict realization of music than of language? Sloboda (1984) remarks that the main purpose of language normally is to convey a message: if the information is understood, the vocalization was good. Since musical meaning is not as obvious, a different demand is placed upon music performance.

2 Theory and hypotheses

2.1 Purpose of this pilot study

This study was designed to render explorative data, by which some aspects of musical sightreading and vocal language reading could be investigated. In the light of previous research, several hypotheses could be formulated as a basis of discussion. A new terminology was also deployed to account for one of the measured entities. New methods of data treatment were tried out in the pilot study.

2.2 Proposal of a new terminology

A deplorable drawback concerning the term “eye-voice span” is that it has little to do with the eyes. One is easily lead to believe that the term denotes the distance between the *visual focus* and the point of performance, when it actually refers to the distance between the subjects visual *attention* and point of vocal performance. Thus, it seems wise to avoid the term eye-voice span and rather talk about the *attention-voice distance* when the entity measured by Levin and others is referred to. Naturally, the distance between the point of fixation and the point of performance is of interest as well, and the term *eye-voice distance* seems appropriate here.

Different methods render different concepts. Eye-tracking, where the eye-voice distance can be measured, has the advantage of really showing where the eyes are directed at any moment. On the other hand, the attention, which directs where

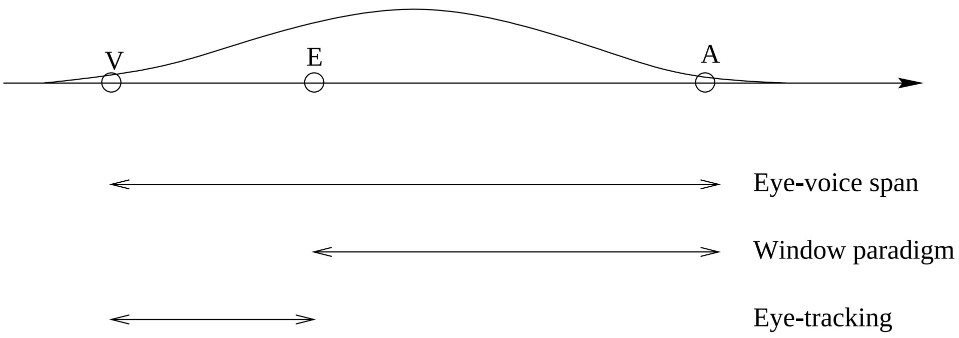


Figure 2: *The relation between the point of vocal performance (V), the eye's point of fixation (E) and the less well defined distribution of attention with its frontal edge point (A) is shown. It is also indicated which distances supposedly can be measured with the three techniques of eye-tracking, window paradigm and eye-voice span.*

information really is apprehended, does not always coincide with the point of fixation.

During the so called eye-voice span experiments (where the stimulus was unexpectedly removed) it is probable that the point of fixation had not reached as far as the last word the subject was able to recall, although attention (with the aid of peripheral vision) had. Thus, these attention-voice distance measurements manage to catch the point of attentional focus, but their limitation lies in interruption of performance and emptying the buffer being unavoidable.

According to the spotlight metaphor, the attention may span a variable range. The relationship between the eye-voice distance and the attention-voice distance is heavily dependent on how much attention diverges from the point of focus. The *attention-eye distance* would be a suitable term for this, and it seems likely that the perceptual spans investigated by the window experiments of Rayner & MacConkie (Rayner, 1978) should correspond closely to this entity.

In figure 2 an attempt is made to put the pieces together. The locations of the voice and eye can easily and unambiguously be measured at any time. The visual attention, however, is a more difficult matter. It appears in the figure as a distribution over an area, and it is assumed that the interesting point in this context is the frontal edge. The tentative assumption is that the attention-voice distance is the sum of the attention-eye distance and the eye-voice distance, and that these entities can be measured with the techniques of eye-voice span (i.e. stimulus removal), window paradigm, and eye-tracking, respectively.

Goolsby (1994b) seems to have confused his own eye-voice distance measurements with assumptions by Sloboda regarding the attention-voice distance. He found the two to be approximately the same, but since no systematic synchronization was made in his study, this result is somewhat obscure. However, the idea to use a time unit in the measurements of perceptual spans deserves to be considered. In fact, it now seems appropriate to differ between six new entities: the *attention-voice spatial distance*, the *eye-voice spatial distance* and the *attention-eye spatial distance* could be measured in words, letters, notes or angular degree, and when they instead take seconds as a unit they can be referred to as the *attention-voice temporal distance*, the *eye-voice temporal distance* and the *attention-eye temporal distance*. Figure 2 accounts for the three spatial distances (in a similar scheme over temporal distances would appear less intuitive, since a mapping between time and space would add to the complexity). In the recordings of this study, the eye-voice temporal distance was measured.

2.3 Hypotheses

Five hypotheses were formulated regarding the eye-voice temporal distance. The average value for music as well as for language was naturally of interest, since the concept is new. The vertical dimension in music (Sloboda, 1985) should make it harder to predict the future saccades, which might decrease the eye-voice temporal distance. Regarding the comparison between text and music, a hypothesis could thus be formulated:

Hypothesis 1: *The eye-voice temporal distance in vocal text reading is generally larger than the eye-voice temporal distance in prima vista singing.*

The possibility that some notes in a piece of music are generally dwelled longer upon than others by the sight-reader should be accounted for. Likewise, in vocal text reading, some words are possibly dwelled longer upon than others. The following hypotheses were stated:

Hypothesis 2a: *The eye-voice temporal distance in vocal text reading is generally longer for some words than for others.*

Hypothesis 2b: *The eye-voice temporal distance in prima vista singing is generally longer for some notes than for others.*

Furthermore, it seemed likely that the eye-voice temporal distance should be affected by the skill level of the subject. Due to the different performance demands

(Sloboda, 1984), musical skill generally differs more obviously between subjects than do their reading proficiency. The following hypothesis was formulated:

Hypothesis 3: *The eye-voice temporal distance of singers differs more between persons while reading music than while reading language.*

The assumption that the average value of the eye-voice temporal distance calculated in this study, along with previous results from window paradigm experiments and the traditional eye-voice span measurements, would fit into the model shown in figure 2 gave rise to the following hypothesis:

Hypothesis 4: *The sum of the average eye-voice distance and the perceptual span of previous window paradigm measurements equals the traditional eye-voice span.*

It was assumed that eye-voice distance would prove a more useful concept in the cases of purely musical or purely textual stimuli than for a stimulus of music with lyrics, since the importance of the vertical dimension in the latter case should make a natural left-to-right flow impossible. However, the music with lyrics may tell us more about the distribution of attention.

Three hypotheses concerning the distribution of attention were formulated. When music with lyrics is read prima vista, attention has to be shared between notes and text. The physical limitations of eye-tracking measurements make the study of this distribution a task fraught with difficulties. Since the resolution is rather poor, the note fixations may be hard to distinguish from the text fixations, at least if the size of the music stimulus is kept within normal limits.

An approach to the study of distribution of attention in prima vista singing that eludes these problems was tried out. The first verse of a tune was as usual written in connection with each music staff, while the second verse had its placement separately below the score. Thus, transitions of attention becomes legible.

The distribution of attention in music with lyrics was to be studied in the case where the lyrics were separated from the score. Since the fixations are generally longer in music reading than in text reading, one could expect that more of the total fixation time should be allotted to the score.

Hypothesis 5: *In music with lyrics separated from the score, the total fixation time on the score is longer than the total fixation time on the lyrics.*

How often will the point of fixation shift between separated music and lyrics, and how many fixations occurs between each such transition? The different reading strategies found in piano score reading suggests that there are at least two possibilities for the eye movement behaviour in the combined stimulus. If the neural buffers are small or dependent on each other, one might expect that only one fixation is made between two transitions. At the other extreme, if the interference between buffers is negligible, and they are of large capacity, a pattern with several fixations on the music may be alternated with a series of fixations on the text. The latter case could indicate that whole sequences of notes and words are needed for an efficient encoding process, while the former one suggests that it is beneficial for buffer storage capacity to combine each note with the right syllable early in the process.

Since the neural evidence is not unambiguous, the fact that music and lyrics phrases coincide was taken as an indication that the combined stimulus was more similar to chordal than contrapuntal music. Hence the transition pattern of the scanpath was expected to resemble the chordal music pattern from the study of Weaver described above, which could be formulated as:

Hypothesis 6: *In music with lyrics separated from the score, scanpath transitions between text and notes are made essentially after every fixation.*

Perhaps speaking against this was a common introspective notion amongst choir singers asked by the author that often a whole phrase is apprehended at a time.

In similar fashion, the transition pattern on music where the lyrics were directly below each staff, was expected to resemble the zigzag pattern that Weaver found in chordal music. In most cases, music looks like this. In order to have an as unambiguous measure as possible of the transition pattern, it was kept in mind that the contrapuntal pattern of Weaver assumes that regressive saccades are made between the two lines. Thus the hypothesis was stated in the following terms:

Hypothesis 7: *In music with lyrics, scanpath transitions between text and notes are seldom regressive.*

The regressions are also of interest in their own right, and they were thus subject to three hypotheses. The findings by Goolsby (1994b) that subjects of different skill level uses regressions for different purposes gave rise to a hypothesis concerning the qualitative nature of regressions in music reading:

Hypothesis 8: *Regressions in vocal music reading may be assigned to different groups depending on the purpose of the eye movement pattern they are a part of.*

The quantitative measurements of Goolsby (1994b) showed that the share of saccades that were regressive was considerably larger in singing than in silent language reading. It was assumed that reading aloud was more similar to silent language reading than to singing, but still resembling singing slightly when compared to silent language reading. This was put to words as follows:

Hypothesis 9a: *The share of regressive saccades is greater in vocal language reading than in silent language reading.*

Hypothesis 9b: *The share of regressive saccades is greater in vocal music reading than in vocal language reading.*

Furthermore, four hypotheses concerning the respective magnitudes fixation duration for the different stimuli were formulated. Fixation duration has been shown to be considerably longer in singing than in silent language reading (Goolsby, 1994a), and it was assumed that the same relation would hold for singing and vocal language reading in this study. The combined stimulus was assumed to induce fixation durations in between of those in language and music reading. Since information is distributed over a larger area when a second verse is separated from the score, the fixations ought to be shorter in that case than when the lyrics is directly below the staffs. Furthermore, since the cognitive load is much greater during second verse singing than during text reading, there was some reason to believe that the fixation duration would be longer in the former case. The following hypotheses were formulated:

Hypothesis 10a: *The fixations are longer during performance of a purely musical stimulus than during the performance of a combined stimulus with music and language.*

Hypothesis 10b: *The fixations are longer during performance of a combined stimulus with music and language than during the performance of a purely textual stimulus.*

Hypothesis 10c: *The fixations are longer during performance of music with lyrics directly below the staffs than during the performance of music with lyrics separated from the score.*

Hypothesis 10d: *The fixations are longer during performance of music with lyrics separated from the score than during the performance of a purely textual stimulus.*

Finally, two hypotheses about the performance durations were formulated. The total time of performance of a stimulus was expected to show more variation in this study than in those where a metronome provided a tempo before each stimulus was presented (e.g. Goolsby 1994a). If the slower tempo is chosen to facilitate reading, one could expect that subjects with poor sightreading ability consequently chooses such a slow tempo over different tunes, whether they have lyrics or not. However, the reading of ordinary text without music will probably rather be performed in a tempo which suits the subjects normal speed of speech. Two hypotheses could thus be formulated:

Hypothesis 11a: *The total performance times of different musical stimuli are correlated across subjects.*

Hypothesis 11b: *The total performance times of musical stimuli are not correlated across subjects with the total performance time of a textual stimulus.*

3 Method

3.1 Subjects

The fifteen participating subjects were all students at Lund University, the age ranging from 18 to 24 years. Most of them were at the time of the recordings members of either of two renowned choirs, Lund Chamber Choir and Lund Student Choral Society, which ensured a reasonably high level of sight reading ability. The subjects could all be classified as skilled musical amateurs (for sightreading studies, a lower level of expertise will result in very inaccurate performance, which is hard to analyze, Goolsby, 1994b).

The selection of participants was based upon availability, hence selection was outside experimental control. The subjects volunteered to participate and were not compensated economically.

3.2 Stimuli

Two old Swedish folk tunes were chosen as musical material. The folk tunes are fairly easy to read prima vista, and not well known to the common chorus singer.

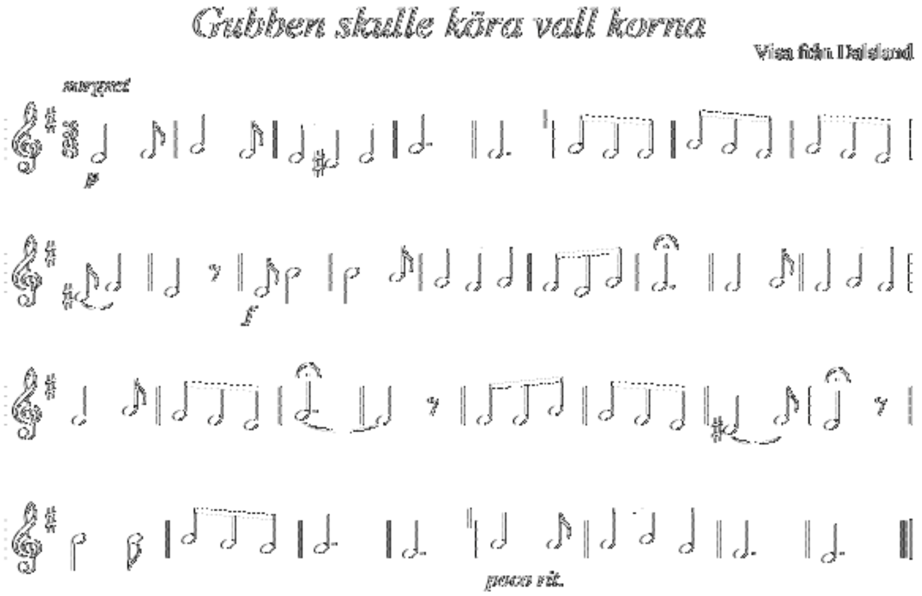
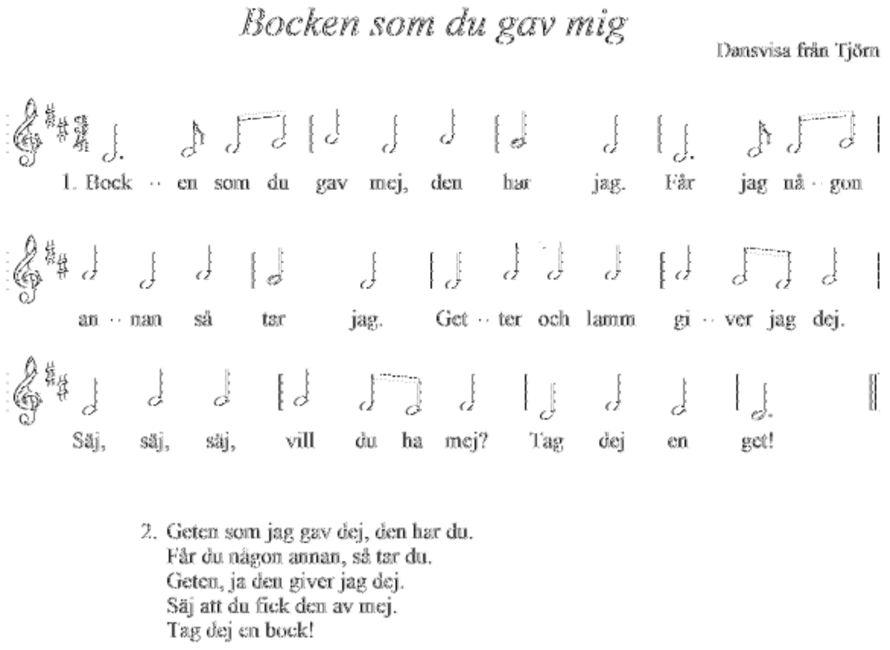


Figure 3: *Swedish folk tune without lyrics.*

The original notes were of poor quality, thus the notes were rewritten in ordinary musical notation (figure 3 and figure 4), by means of a music notation computer program. Both pieces were notated in treble clef, and both melodies started with their keynotes. The ranges were small, a minor sixth and a perfect fifth, respectively.

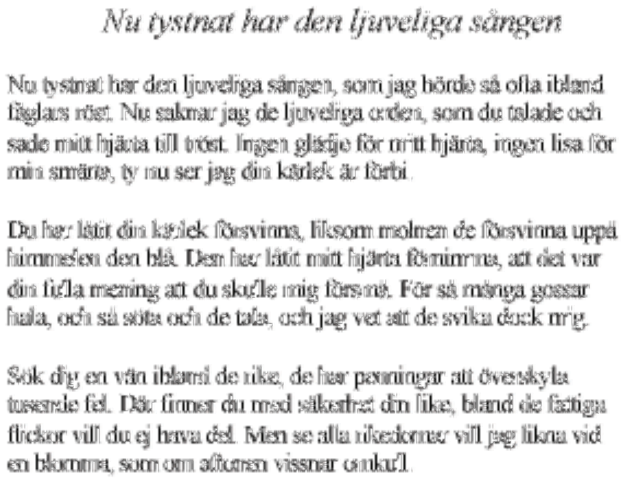
Several modifications were made to achieve the tune in figure 3: the text was removed, as well as two accompanying parts. Furthermore, the original key was lowered a fourth into E minor, in order to facilitate for the lower voiced male subjects. Breath marks and fermata, as well as markings of tempo and dynamics, were not altered or added to. The mood instruction “sorgset” is Swedish for “with sorrow” and “Visa från Dalsland” means “Tune from Dalsland”. The tune consisted of four staffs, in all 68 notes, scattered across 33 bars of 3/8 meter.

The tune of figure 4 was already written in a comfortable D major. The first verse of the text was written beneath the corresponding notes, while the second verse was placed by itself below the musical notes, a format well known from e.g. Swedish church psalms. The original music contained no dynamical or temporal markings, and none were added. “Dansvisa från Tjörn” means “Dance tune from Tjorn”. The folk tune was written in 3/4 meter, and its 12 bars consisted of 37


Bocken som du gav mig
Dansvisa från Tjörn
1. Bock - en som du gav mej, den har jag. Får jag nå - gon
an - nan så tar jag. Get - ter och lamm gi - ver jag dej.
Säj, säj, säj, vill du ha mej? Tag dej en get!

2. Geten som jag gav dej, den har du.
Får du någon annan, så tar du.
Geten, ja den giver jag dej.
Säj att du fick den av mej.
Tag dej en bock!

Figure 4: *Swedish folk tune with lyrics.*



Nu tystnat har den ljuveliga sången

Nu tystnat har den ljuveliga sången, som jag hörde så ofta ibland
fåglars röst. Nu saknar jag de ljuveliga orden, som du talade och
sade mitt hjärta till tröst. Ingen glädje för mitt hjärta, ingen lisa för
min smärta, ty nu ser jag din kärlek är förbi.

Du har låtit din kärlek försvinna, liksom molnen de försvinna uppå
himmelen den blå. Den har låtit mitt hjärta förnimma, att det var
din fulla mening att du skulle mig försmå. För så många gossar
hala, och så söta och de tala, och jag vet att de svika dock mig.

Sök dig en vän ibland de rike, de har penningar att överskyla
tusende fel. Där finner du med säkerhet din like, bland de fattiga
flickor vill du ej hava del. Men se alla rikedomar vill jag likna vid
en blomma, som om aftonen vissnar omkull.

Figure 5: *Lyrics from a Swedish folk tune, arranged as ordinary text.*

notes and stretched across tree staffs.

Finally, a text was taken from a third folk tune[2] (figure 5). Three verses were written as separated units, but the line breaks did not follow the natural rhyme and phrase pattern of the song, since this is not the case in musical notation, with which comparisons were to be made. The total number of words was 147.

Coarse translations of the textual stimuli are found in appendix C.

3.3 Apparatus

The recordings were made at LUCS Eye Tracking Laboratory in Lund, Sweden. An SMI iView remote eye-tracker device was set up below the paper sheets of music and text. The music notation was copied with a slight magnification onto sheets that were attached in front of each other, in a note pad manner. The subject was seated with an approximate distance of one meter between eyes and the sheets to be read, and the notes probably appeared somewhat larger than usual.

[2] The language of the purely textual stimuli was somewhat poetic. A few uncommon words and word forms differed from everyday Swedish, and it is possible that the exact meaning of all the words was not known to every subject. The atmosphere in the performance of this text is similar to the one in singing performance. The rhymes and the poetic hue of the words encourage the reader to put a little more effort to the performance quality. The reason to choose such a text was that there should be similarities with the lyrics of the stimulus containing both music and language. Even though the themes of the texts are not the same, similarities, such as rhyming, clear phrase boundaries and the performance atmosphere already mentioned, provides a good basis for comparisons.

The eye-tracker uses an infrared reflection in one of the subjects corneae to record the gaze of the subject at 50 Hz. The equipment thus enables to spot every fixation with an accuracy of one angular degree.

The eye movement data were complemented with a video camera placed close to the head of the subject, which by aid of a video editing program provided a means of matching each vocal expression to its corresponding eye position with an accuracy of 40 ms.

3.4 Procedures

Three successive recordings were made on each subject. The procedures were rehearsed with each subject using a musical note of an additional Swedish folk tune, without recording. In a short procedure to calibrate the eye tracker, the subject was told to sit comfortably but straight, and to look at nine glowing spots in a certain order.

First, a musical note without text was presented, second, a note with two verses of text. Both were to be sung prima vista. Finally, an ordinary text without notes was to be read aloud.

Instructions were given just before each recording. The subjects were told that they were allowed to look at the notes a few seconds as the paper hiding it was removed, but encouraged to start singing as soon as possible (using the syllables "noh, noh, noh ... " on the music without lyrics). First they should sing the tune by themselves, and then a second time when the author would sing along. The subjects had to imitate a short phrase consisting of the first five notes of the key to come, and the first note of the tune was sung to them.

When mistakes followed by hesitation were made during recordings where the subjects were to sing alone, the author sang along as support until the subject was back on the correct melody.

Before the recording of ordinary text reading, the subjects were simply told to read the text aloud.

4 Data treatment

4.1 Modifications and data loss

Some subjects altered their body position slightly between calibration and singing procedures, which resulted in an offset in the eye movement data. The subjects were assuming a more tense and upward position during calibration, compared to the more relaxed and forward position which enables better breathing during singing. However, the body position was seemingly constant during the actual recordings, and the offset was assumed to be constant.

Another problem that occurs in eye-tracking is that data sometimes is lost when the camera is unable to get a clear view of the light reflected from the eye. Data was thus incomplete from some recordings, due to odd reflection in soft contact lenses, blocking by glasses frames, and excessive head movement. The number of subjects used in the statistics varies, but nine data sets were good enough to be used in all the calculations.

4.2 Spatial objects

In the music without lyrics, two spatial objects were defined: the music and the lyrics of the first verse constituted one object, the lyrics of the second verse the other, as in figure 6. These will be referred to as the music object and the text object, respectively. The eye-tracking software admitted a display of distribution of fixations as a function of time, with regard to the two spatial objects. The lower time limit of what should be counted as a fixation was set to 100 ms.

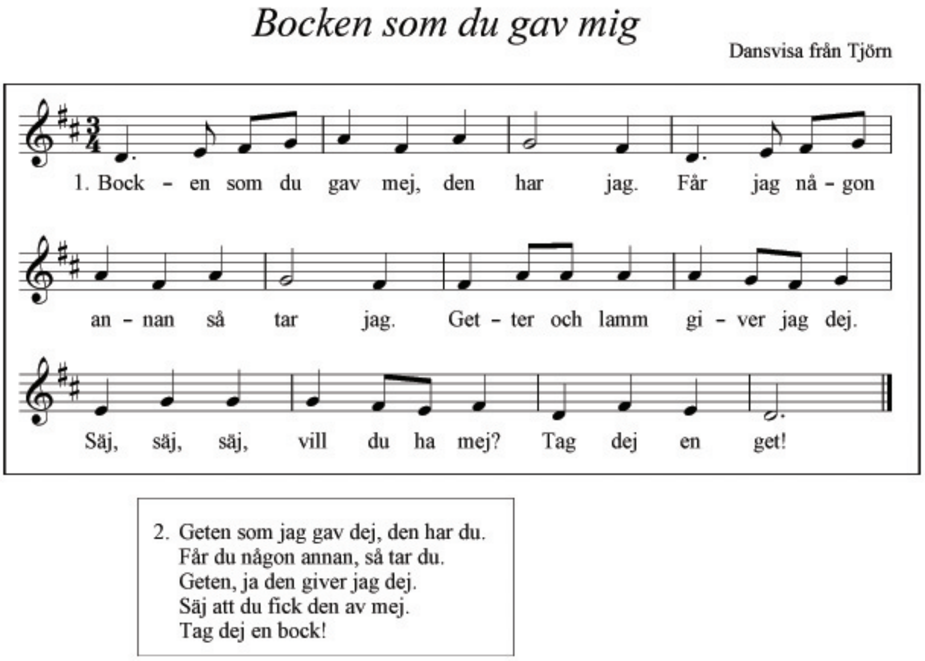
4.3 Regression filter

In order to measure the number of regressions, a simple filter was created in MS Excel, where fixations to the left of the previous fixation was counted. The Excel formula can be seen in figure 7.

Return sweeps were ignored, as well as fixation shifts which had a large vertical component, so that almost vertical saccades would not be confused with regressions. In addition, a short regression was selected as a minimal one, in order to avoid the situation where a fixation very close to the previous one (on the same note head or letter) but slightly to the left of it is counted as a regression. The filter did not detect the regressions that were made across line breaks, but these were rather few, and are accounted for elsewhere.

4.4 Eye-voice temporal distance

The eye-voice span has traditionally been measured as a spatial range. However, it is not evident that the distance on the music sheet is more appropriate a dimension than time of performance (Goolsby, 1994b). A temporal measure of the eye-voice distance was tried out in the data treatment of this study. However, there was no


Bocken som du gav mig
Dansvisa från Tjörn
1. Bock - en som du gav mej, den har jag. Får jag nå - gon
an - nan så tar jag. Get - ter och lamm gi - ver jag dej.
Säj, säj, säj, vill du ha mej? Tag dej en get!

2. Geten som jag gav dej, den har du.
Får du någon annan, så tar du.
Geten, ja den giver jag dej.
Säj att du fick den av mej.
Tag dej en bock!

Figure 6: *The rectangular areas each constitutes an object. The larger, upper rectangle will be referred to as the* music object, *while the smaller, lower one will be referred to as the* text object.

```
=IF(AND((E1-E2>14),(E1-E2<300),(ABS(F1-F2)<14)),1,0)
```

Figure 7: *Simple regression filter in MS Excel. The x-coordinates of the fixations were stored in column* `E`*, while the y-coordinates were stored in column* `F`. `E1-E2` *thus stands for the horizontal distance between the first and the second fixation, while* `ABS(F1-F2)` *is the vertical distance (the* `ABS` *operator takes the absolute value). The* `AND` *operator links three conditions and makes sure that the* `IF` *operator tests for all tree requirements to be fulfilled: to count as regressive, a fixation must be at least 14 and not more than 300 points to the left of the former one, and it may not diverge more than 14 points in the vertical direction. The formula results in 1 if a regression is detected, 0 otherwise. In MS Excel the reference numbers are modified automatically when the formula is copied to the next row, so that e.g.* `E1-E2` *is transformed into* `E2-E3`*. Addition of the result from each row results in the total number of regressions.*

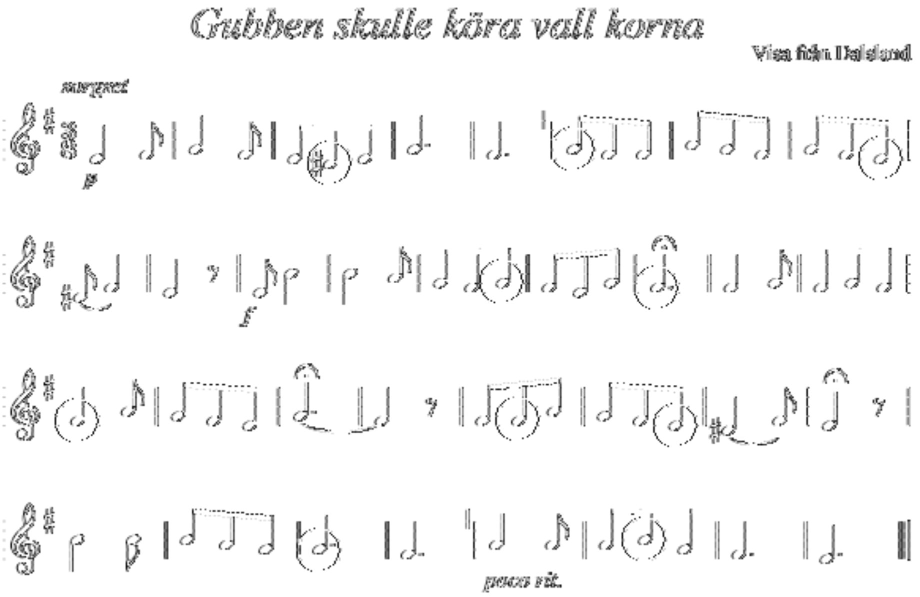


Figure 8: *Ten notes selected as places to measure the eye-voice temporal distance.*

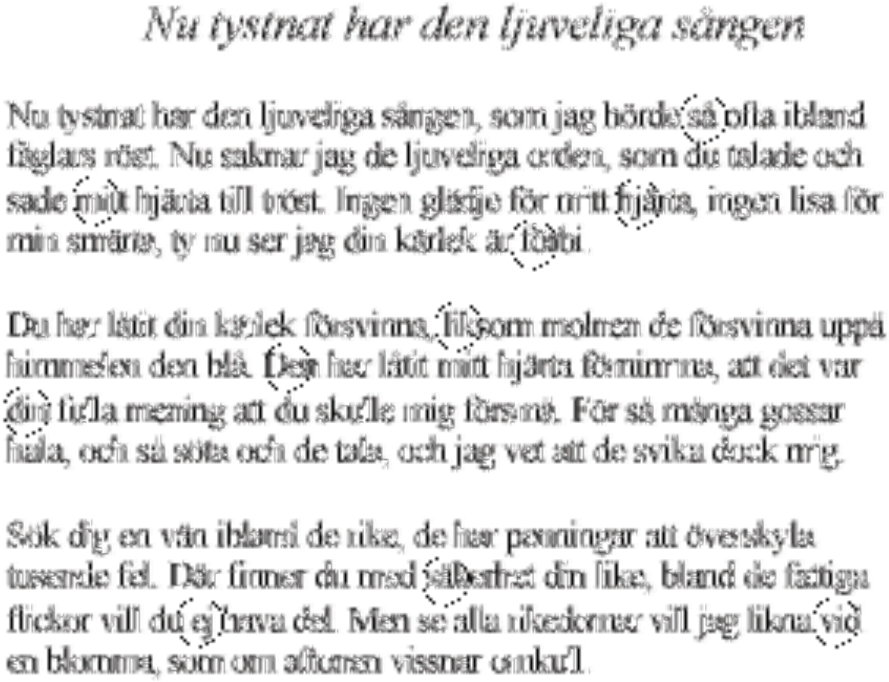
possibility to elicit the eye-voice temporal distance as a continuous function of time. Instead, ten notes were chosen in the purely musical stimuli, as well as ten words in the purely textual stimuli, and the eye-hand span was measured at these loci.

The notes were chosen to be of different distances from phrase shifts, line breaks, large intervals, bar lines, and emphasized notes. The notes are encircled in figure 8. The words were chosen to be of different distances from phrase shifts, line breaks, uncommon words and emphasized words. They are encircled in figure 9

The eye movement and video recordings was studied through a video editing program. When the subject saccades reached a selected note, the number of 40 ms steps until the corresponding vocalization begun was extracted and taken as a measurement of the eye-voice span.

5 Results

The hypotheses in section 2.3 will be referred to frequently below.



Nu tystnat har den ljuveliga sången

Nu tystnat har den ljuveliga sången, som jag hörde så ofta ibland fåglars röst. Nu saknar jag de ljuveliga orden, som du talade och sade mitt hjärta till tröst. Ingen glädje för mitt hjärta, ingen lisa för min smärta, ty nu ser jag din kärlek är förbi.

Du har låtit din kärlek försvinna, liksom molnen de försvinna uppå himmelen den blå. Den har låtit mitt hjärta förnimma, att det var din fulla mening att du skulle mig försmå. För så många gossar hala, och så söta och de tala, och jag vet att de svika dock mig.

Sök dig en vän ibland de rike, de har penningar att överskyla tusende fel. Där finner du med säkerhet din like, bland de fattiga flickor vill du ej hava del. Men se alla rikedomar vill jag likna vid en blomma, som om aftonen vissnar omkull.

Figure 9: *Ten words selected as places to measure the eye-voice temporal distance.*

5.1 Eye-voice temporal distance

On average, the eye-voice temporal distance was 500 ms when music without lyrics was performed, and 750 ms when text without music was performed. The standard deviations were 250 ms and 560 ms, respectively. A two-sided t-test showed that eye-voice temporal distance in vocal language reading was generally larger than in prima vista singing ($p = 0.011$), which positively confirmed *hypothesis 1*.

The eye-voice temporal distance was, according to an ANOVA with a Tukey HSD *post hoc*-test, not significantly larger across the subject population for any specific note. *Hypothesis 2a* could thus not be verified with the limited amount of data at hand.

However, in the vocal language reading, the eye-voice temporal distance at the fifth as well as at the sixth word encircled in figure 9 was longer than at the first one, which confirms *Hypothesis 2b*. An ANOVA with a Tukey HSD resulted in a significance level of $p < 0.05$ for both comparisons. The words of long eye-voice temporal distance were both at the beginning of a phrase.

The total singing time of the first tune correlated to the eye-voice temporal distance ($r = 0.57$, $p < 0.05$). The total reading time of the text stimulus showed no such tendency.

The standard deviations for the eye-voice temporal distance in music was compared to those in language with a t-test, which rendered no significance. The variation between subjects was thus not larger in music reading and *hypothesis 3* was not verified.

Even though no measurements were made for the eye-voice spatial distance, a rough estimate of this was calculated by dividing the temporal eye-voice temporal distance with the total time and multiplying it with the total number of notes. The result was an eye-voice spatial distance of 0.6 notes, which seemed to be reasonable in an inspection of the video recordings. Likewise, the eye-voice spatial distance in text reading was estimated to 8.9 character positions, equalling 2.0 words. Previous research on the eye-voice span in language reading has shown that the attention-voice spatial distance is 5–6 words (Sloboda, 1984), while the window paradigm experiments have resulted in an attention-eye spatial distance of 10-15 letters (Rayner, 1978). If these character positions equals 2-4 words, the assumptions illustrated in figure 2 seems promisingly good. *Hypothesis 4* cannot be rejected with the evidence at hand.

5.2 Distribution of attention in music with lyrics

The attention during the second verse of the song with lyrics was almost equally shared between the music object and the text object in figure 6. The value of attention distribution was on average 52% on the music object and 48% on the lyrics object, with a standard deviation of 9%. A t-test showed that the difference was not significant, and *hypothesis 5* was thus falsified, since the total fixation time was not longer on the music object than on the text object.

It should be taken into account that some of the fixations within the music object was on the lyrics of the first verse. Unfortunately, the data was generally too ambiguous to make a quantitative analysis. Only one of the data sets was accurate enough for the distribution of attention between music and lyrics in the music object to be explored more carefully. That particular one revealed that 34 per cent of the time spent on fixations in the music object the eyes were directed at the lyrics of the first verse during the performance of the second verse.

In order to explore a situation more similar to normal choir singing, the distribution of attention during the performance where the author and the subject sang in unison was measured as well. The result differed from the solo performance: 40% of the fixation time was spent within the music object, and 60% on the second verse lyrics. However, the standard deviation was considerably larger (27%). Surprisingly, one subject looked at the text object only.

Between the spatial objects, defined in figure 6, 1.48 transitions occurred every second, on average. The transition rate can also be expressed in the following terms: on average, 2.9 fixations are made between every transition, with standard deviation of 0.9 fixations. This implies that the total fixation time on one type of stimulus, before a transition occurs, is three times as large as the mean fixation time, 735 ms to be precise. *Hypothesis 6* is falsified by this result. Apparently there are short sequences of fixations between transitions, an example of which is shown in figure 10. However, the case where only one fixation occurred between two transitions was also seen, which is exemplified in figure 11. The transition saccades were directed at the point of performance and not ahead of it.

The number of transitions was tested for correlation with the total performance time of the second verse. There was a tendency of correlation, however not significant. One of the subjects appeared to have rather deviant values compared to the others, in this respect as well as in others (*inter alia* he was the one who made three verse mix-up errors). If this person is taken as a statistical outlier, the correlation is of clear significance ($r = 0.80$, $p < 0.01$).

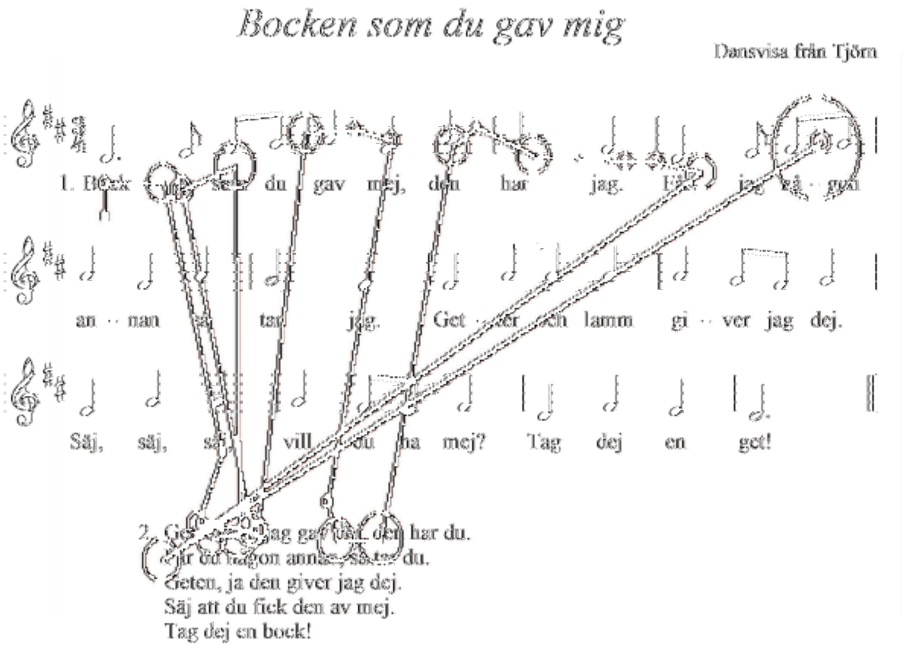

Bocken som du gav mig
Dansvisa från Tjörn
1. Bock - en som du gav mej, den har jag. Får jag så - gun
an - nan så tar jag. Get - ter och lamm gi - ver jag dej.
Säj, säj, säj vill du ha mej? Tag dej en get!
2. Geten som jag gav dej, den har du.
Får du någon annan, sa ter du.
Geten, ja den giver jag dej.
Säj att du fick den av mej.
Tag dej en bock!

Figure 10: *Scan path with a "contrapuntal" transition pattern. The scanpath of this figure was fitted to the music with a separate image manipulator program, hence the distorted circles and unprecise path.*

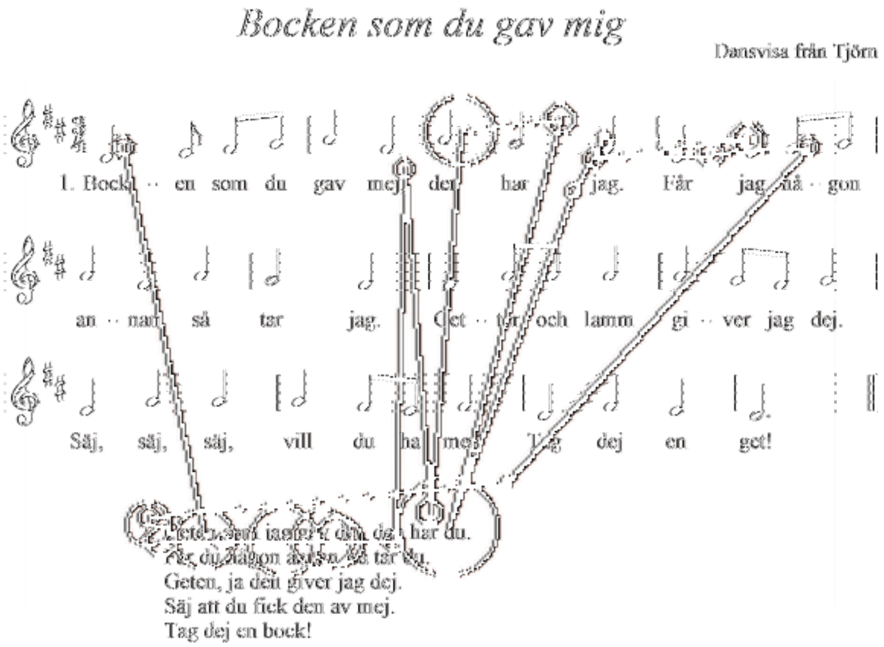

Bocken som du gav mig
Dansvisa från Tjörn
1. Bock - en som du gav mej den har jag. Får jag nå - gon
an - nan så tar jag. Get - ter och lamm gi - ver jag dej.
Säj, säj, säj, vill du ha mej? Tag dej en get!
Får du någon annan så tar du.
Geten, ja den giver jag dej.
Säj att du fick den av mej.
Tag dej en bock!

Figure 11: *Scan path with a partially“chordal” transition pattern. The scanpath of this figure was fitted to the music with a separate image manipulator program, hence the distorted circles and unprecise path.*

5.3 Regressions

Several different kinds of regressions were separated in the analysis, and *hypothesis 8* should be considered true.

The lion's share of the regressions occurring in sightreading displayed the same pattern: the subject looked further ahead in the music, then regressed to the point of performance, the eyes and voice taking the next step simultaneously. In other words, the eye-voice temporal distance is reduced to zero in the regression of this saccade pattern. These regressions were never extended across more than one interval.

Some of the regressions derived from performance errors; the subject looked back to the place where the error occurred in order to find the correct pitch. Some regressions were directed at notes which were already performed. Four of these targeted the first note of a staff, and crossed several notes on their way. The target notes were at the first and the third staff.

As much as ten regressions were executed at phrase beginnings, and at least three of these actually took place at the same time as — or even after — the first note of the new phrase was articulated.

Four regressions were made across line breaks in the music, and they occurred before the first note on the new line was articulated. One regression was aimed at an accidental at the beginning of the line.

Only a few regressions resulted from the subject looking ahead in the lyrics, and returning to the point of performance in the notes and vice versa. The lyrics was instead apprehended by dips in the sequence of progressive saccades, and *hypothesis 7*, was thereby verified.

The results of the quantitative measurement of regressions are shown in table 1. Since there were much variation between subjects, the standard deviations are rather large. Taking this into account, the share of regressions in vocal language reading, should lie somewhere between 1.0 and 8.6 per cent, and in music reading between 3.6 and 11.2 per cent. *Hypothesis 9a* was formulated to compare vocal and silent language reading. However, since the result differs considerably from previous research in both music (where the share was 30%, Goolsby, 1994b), and language (where the share was 10–15%, Reichle et al., 2000), it seemed wise to discuss this result before any conclusions were drawn.

The differences between stimuli presented in table 1 were not significant when a t-test was applied. However, although *hypothesis 9b* could not be verified, the tendency was in the predicted direction, i.e. that more regressions occur in music

Table 1: *The share of regressive saccades in per cent.*

stimulus	music without lyrics	music with lyrics		text without music
		verse 1	verse 2	
regression share	7.4	6.8	7.7	4.8
standard deviation	3.8	4.4	3.8	3.8

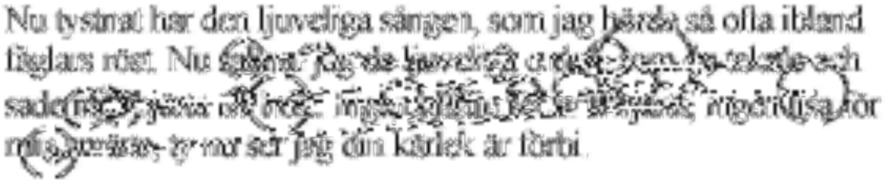

Nu tystnat har den ljuveliga sången, som jag hörde så ofta ibland
fåglars röst. Nu
nu ser jag din kärlek är förbi.

Figure 12: *A typical scanpath in text reading. The scanpath of this figure was fitted to the music with a separate image manipulator program, hence the distorted circles and unprecise path.*

sightreading than during vocal language reading.

5.4 Vertical dimension

In the ordinary text the vertical dimension was important at line breaks only. The scan paths followed each row of text without much vertical deviation. Figure 12 shows a typical such scan path. In music, the vertical dimension manifested itself more. The scan paths approximately traced the contour of the music, and large intervals were reflected as distinct saccades that diverged from the horizontal dimension at a high degree. This was shown in figure 1.

Some completely vertical saccades occurred. Five of those were at fermatas, two at the fifth in bar eleven of the music without lyrics, and eight between text and music in the first verse of the tune with lyrics.

In the second verse of the music with lyrics the necessity of large vertical saccades was commented as strenuous.

5.5 Fixations

The mean fixation times are shown in table 2. An ANOVA with a Tukey HSD showed that the differences were significant ($p < 0.01$) between all of these average values except for between those of the text and the second verse of the music with lyrics. This result was precisely as predicted in *hypotheses 10a–c*, but the

Table 2: *Mean fixation times in milliseconds.*

stimulus	music without lyrics	music with lyrics verse 1	 verse 2	text without music
mean fixation time	465	370	249	248
standard deviation	70	60	43	21

more vague assumptions that served as a basis for *hypothesis 10d* did apparently not tell the whole truth, and that one was not verified by the results.

In search for individual differences, correlations were found between music without text and music with text ($r = 0.88$, $p < 0.01$ for verse 1; $r = 0.68$, $p < 0.05$ for verse 2), as well as between the two verses of the music with lyrics ($r = 0.76$, $p < 0.05$). However, no correlations were found between the purely textual stimulus and any of the musical ones.

On average, the subjects made 1.6 fixations per note on the purely musical stimulus, and 1.1 fixations per word on the purely textual stimulus. The standard deviations were 0.3 and 0.1 fixations, respectively.

5.6 Tempo difference

Since no tempo was provided the subjects were free to choose a tempo of their own. The average values of the total time of performance varied as shown in table 3.

In search for individual tempo preferences, a correlation was found between the total performance time of the tune without lyrics and the total performance time of the vocal text reading ($r = 0.80$, $p < 0.01$). Also, significant correlations were found between the total performance time of the second verse of the tune with lyrics and the total performance time of the text reading ($r = 0.74$, $p < 0.01$), as well as between the total performance time of the second verse of the tune with lyrics and the total performance time of the tune without lyrics ($r = 0.68$, $p = 0.010$). This entails that *hypothesis 11a* is confirmed, since the total performance times of different musical stimuli do correlate. However, *hypothesis 11b* was not true, since the total performance time of the textual stimulus correlated to the musical performance times.

5.7 Performance errors

Many errors were made during the performances. The highest error frequency occurred during the prima vista singing of the second verse of the tune with lyrics.

Table 3: *Total time of performance in seconds.*

stimulus	music without lyrics	music with verse 2	text without music
time of performance	38	60	54
standard deviation	7	11	6

Several errors also occurred during the first verse of that piece, as well as during the performance of the tune without lyrics. Only a few mistakes were made in the vocal text reading. When the author and the subject sang together, there were almost no errors at all. A further investigation of the performance errors provides clues to the interpretation of the eye-tracking data. A listing of errors might also prove useful as an aid to understanding the processes of sightreading on the whole.

Many subjects made interval errors, and often they noticed it themselves. The immediate action was usually to back up in the performance and try to correct the error. However, six times a performed note was revocalized even though it was correct the first time. This also happened twice in the vocal text reading. Most interval errors were made at large intervals. A common place for these errors was the first interval of each song that was larger than a second. Four persons made mode mix-ups, i.e. performed minor where major was due or vice versa.

As much as eighteen interval errors resulted in a whole series of subsequent notes being performed at an incorrect pitch, one note above or one note below the actual melody line. During these passages, the absolute intervals were often incorrect, since the subjects kept the performed notes within the key (correct intervals would sometimes have resulted in notes not fitting into the key). Thus, the melodies which were performed were not actual transpositions, but so called *tonal answers*. The tonal answer errors lasted between one and fifteen bars. They ended either by a performance of a correct interval that resulted in a note out of key (which often disrupted the performance, cf. Banton, 1995), or by the subject simply singing the correct pitch of one note, seemingly unaware of any correction being made, and continuing without errors.

An error occurring in performances of the tune with lyrics, was that a word from the lyrics of the first verse was sung during the performance of the second verse. In all, six such errors were made, three of them by one subject. This was the subject of whom the attention distribution within the music object was studied. Furthermore, three subjects made the error of pronouncing the word “dej” as “dig”.

There are two pronunciations of this word, which means “you” (in the singular form) in Swedish[3], but no one would pronounce a written “dej” as “dig” under normal reading conditions, when there is time to think.

Very few rhythmical errors were made, i.e. if the pauses of hesitation are disregarded. There was a lot of those, though.

6 Discussion

6.1 Eye-voice temporal distance

The eye-voice temporal distance was found to be significantly larger in language reading than in music reading. An explanation that seems plausible is that, due to the greater importance of the vertical dimension, it is harder to predict the direction of eye movements in music reading. Too much eye-voice delay would result in less precise control, and probably more vertical and less efficient progressive saccades.

The standard deviations of the eye-voice temporal distance did not differ significantly between language and music. This entails that the musically skilled subjects did not prove to have a more stable eye-voice distance than the poorly skilled.

The eye-voice temporal distance was not significantly larger for any specific note, and if such differences exist, a larger amount of data has to be analyzed to detect them. Presumably, the phrase shifts are of special interest. Partly so because previous research has shown that the attention-voice spatial distance is affected at these points (Sloboda, 1984), and, since there probably exists some sort of relationship between the attention-voice spatial distance and the eye-voice temporal distance, the phrase boundaries are likely to influence them both.

Another reason to aim at the phrase boundaries in the exploration of the musical eye-voice temporal distance are the significant differences between words in the text. The longest eye-voice temporal distances were at the first words of phrases. This result is possibly due to the small pauses that subjects insert between phrases

[3] In Swedish, this personal pronoun is usually spelled “dig”. The normal pronunciation, however, would undisputedly be spelled out “dej”, according to the standard relationship between Swedish phonemes and graphemes. Fredrik Lindström explains in his book *Världens dåligaste språk* that historically, people who ambitioned belonging to the upper class believed that the most distinguished spoken language was as similar to the written language as possible (even though the real upper class knew the difference between speech and writing). Thus, a common notion is that a literal pronunciation of “dig” is more correct than the actual pronunciation “dej”. Choirs who wish to convey a solemn and serene atmosphere, often use the former one. What complicates this case is that the normal speech pronunciation is spelled out in the lyrics. This is commonly viewed upon as a very informal spelling, but since folk tune lyrics often are closely tied to dialectic pronunciation the spelling is not entirely improper in this context.

when a text is read aloud. During these pauses, there is additional time for the gaze to dwell upon the first word of the next phrase. Nevertheless, a speculation about the relationship between attention and point of fixation could be of interest. In text reading, more information is normally collected to the right of the point of fixation than to its left (Sloboda, 1978). Thus, it is reasonable to presume that the center of the attention spotlight is ahead of the point of fixation. It is also known that the attention-voice spatial distance is increased when a phrase ending is just beyond the average span (Rayner, 1978), and then decreased as the phrase ending is approached. There is a possibility that at the end of a phrase, the lead of the attention over the point of fixation is decreased, since attention takes some time at the last words of the phrase. When the point of fixation gets too far to the right in the attention spotlight, which would be at the first word of the next sentence, the eye movements are slowed down until the attention catches up.

The question of comparable units in music and language has no easy answer. Words share similarities with melodic features such as arpeggios and scale movements, but also with rhythmical patterns and, in notation, with the single notes. The selected words in the textual stimulus was supposed to correspond to the selected notes in the music stimulus. One may argue that, in language, perhaps the closest correlate to a musical note is a syllable — or even a phoneme — rather than a word. Thus, an option would have been to include some syllables that were not in the beginning of a word in the selection. On the other hand, there is white space between all note heads, but not between syllables of the same word, i.e. visually, the connection may be closer between words and notes. Then again, when note tails are joined, the notes are more likely to be understood as a unit. No matter the choice of unit, it was evident from the data that the results would not have differed much.[4]

6.2 Relations between distances

The eye-voice spatial distance in vocal language reading was approximately two words. In figure 13, an attempt is made to explain the relationship between this visual distance and two other ones. The attention-voice spatial distance (5–6 words, from eye-voice span experiments, Sloboda, 1984) seems to be the sum of the eye-voice spatial distance (2 words, from the present eye-tracking study) and the

[4] Chafe (1994) demonstrates how parallels may be drawn between intonation units in language and music, and perhaps the more detailed level of words and notes is unsuitable if a deeper understanding of the similarities between music and language is sought.

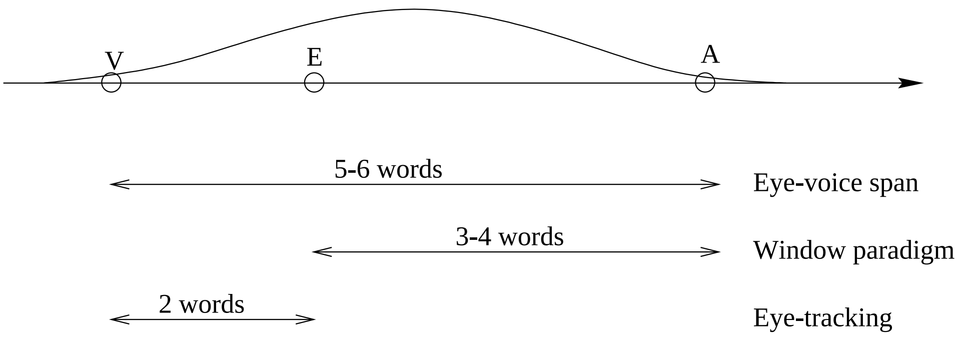


Figure 13: *The relation between the point of vocal performance (V), the eye's point of fixation (E) and visual attention (A) is shown (the curve symbolizes the actual distribution of attention, where point A had reached the furthest ahead in the text). Three distances, that supposedly can be measured with the three techniques of eye-tracking, window paradigm and eye-voice span, are represented.*

attention-eye spatial distance (2–4 words, from the eye-voice span experiments, Rayner, 1987). Although the point of attention is vaguely defined, the number of words fit together strikingly well.

The copy typists (Sloboda, 1985) who only needed eight character positions to maintain speed, can be compared to the subjects in a window experiment where more letters affect performance. It seems reasonable that in a manual copying task there may be constraints not present in the silent reading of the window experiment. Thus, the window paradigm seems to be the better choice for this relation discussion. Optimally, though, results from vocal language reading window experiments would have been desirable, since it is a little awkward to compare vocal distances with silent ones. The relation model cannot yet be verified for note reading as well, since there is no data on vocal reading in a window experiment.

6.3 Regressions

Previous research states that 10–15% of the saccades in silent language reading are regressive (Reichle et al., 2000). Even though the variation of the share of regressions was large between subjects, the numbers in the present study of vocal language reading are clearly lower than this, viz. between 1.0 and 8.6 per cent (table 1). It seems puzzling that fewer regressions are made during vocal reading than during silent reading, since the silent reading is faster and there should be less time to move your eyes back and forth.

However, the difference from previous research is even more striking for music reading: Goolsby (1994a) had a 30 per cent share of regressions, but in the present study the share was between 3.6 and 11.2 per cent. It is possible that the lower value of regression share do not reflect a difference between vocal and silent reading, or between different musical stimuli, but rather differences in the definitions of a regression. Goolsby (ibid.) states that all the right-to-left eye movement are regressive. However, there should be some limit on how short the saccade is allowed to be at a minimum, lest the tiny eye movements within fixations may count as regressive saccades. Furthermore, Goolsby (1994b) sometimes counts saccades directed at the lines not yet performed as regressive, when there clearly must be some difference between look-backs at the already performed parts and look-aheads at the notes not yet performed.

Apparently, comparisons between the regression share results of this and the previous studies have poor validity. A common definition of regression is needed, which has to be more realistic than the definition of Goolsby (1994a) and more precise than the definition used in the present study.

There is a tendency in this study that there are more regressions in music than in language. The complexity of the stimuli may be of importance, and in this case the music was fairly simple while the text was fairly complex. The tendency is also valid if the regression results in Reichle et al. (2000) are compared with those in Goolsby (1994b).

Most regressions in sightreading reduced the eye-voice temporal distance to zero. This was the pattern that Goolsby (1994b) found amongst good readers, and the frequency was high enough to suspect that poor readers make use of this pattern too. The importance of this strategy becomes evident when the fact that it even occured across line breaks is considered.

It seems that subjects needed to refresh their memory just before the new note was performed, in these tunes most likely the memory of pitch, since the rhythmical patterns were simple. Either the subjects had already observed the interval, but lost it due to some interference effect, or the intervals were never apprehended during the first glance so that a regression was necessary to fill out the missing information. A third possibility is that the subjects knew the right interval, but just wanted to check, for safety’s sake. The errors where subjects revocalized despite their first attempt being correct makes it plausible that subjects are uncertain of the correctness of their performance also at other occasions than those leading to performance breakdown.

This uncertainty could possibly also account for the regressions at phrase beginnings where the first note of the phrase already was articulated when the a regression to the previous note occurred.

The regressions that were aimed at the first none of a staff seems peculiar. The note at the beginning of the first staff is the keynote, and the note at the beginning of the third staff is fundamental note of that particular phrase. An explanation not too far-fetched is that the subjects remembered the pitch of the first note of the staff, but had to check its vertical position in order to compare it with the upcoming note. Deutsch (1970) concluded that the interference effects are large even when the intervening notes are tonal, but it seems that in some cases a previous absolute pitch can be remembered, at least if it has an important role in the current melody lines.

6.4 Distribution of attention in music with lyrics

The distribution of attention between the score and the second verse of the tune in figure 4 was almost equal. However, when the author sang along, the distribution altered in favour of the text object, in one case nothing but the text object was fixated. It is difficult to say whether this was due to the subject learning the melody (this was the fourth time the subject sang the melody of the tune) or if the subject relied on the authors song as an aid for pitch rather than an aid for words. The standard deviation was larger when the researcher sang along, possibly since there was probably more room for individual preferences in the distribution of attention when the task was made easier.

The task of reading the second verse of a tune, when the verse was written below the score, was thought of as strenuous. Thus, the point of eye focus ought to have been moved as efficiently as possible and it is unlikely that the attention would be directed elsewhere. In the light of these thoughts, conclusions for attention can be drawn, with some certainty, directly from eye movement data.

The overall transition pattern resembled chiefly that of contrapuntal piano music (Sloboda, 1985). However, the pattern of the first verse was rather of the chordal type, which suggests that attention is distributed fundamentally differently in the two cases. This makes it difficult to draw comparative conclusions from the two verse performances.

The total fixation time between transitions during the second verse was approximated to 735 ms. Assuming that information is taken in during all fixation time, this time value is the amount of information stored in the buffer before each transition. Since the distribution of attention was fairly equal, 735 ms should be a good approximation for both the music buffer and the language buffer[5]. But is this the maximum storage capability?

An attention shift might occur whenever a buffer becomes empty, which seems to be a good assumption if the tempo is hard to keep up with. If the attention is well ahead of performance, it may instead be executed whenever a buffer is filled up. The transition saccades were directed at the point of performance, thus it seemed that the criterion for the execution of a transition saccade was that the other buffer became empty, not that the current buffer was getting full. Thus, no conclusions can be drawn about the maximum storage capability, and 735 ms merely indicates how much ahead of performance the attention was (i.e. the attention-voice temporal span) at each time of transition.

The correlation between the number of transitions and the performance time entails that each subject stored an approximately equal amount of temporal information between transitions. This indicates that milliseconds would be a better buffer unit than the number of notes.

The results from the distribution of attention within the music object should merely be considered as tendential, considering that eye-tracking accuracy is comparatively poor when the head is not stabilized. The subject who made three verse mix-up errors spent one third of his fixation time within the music object looking at the lyrics of the wrong verse. This seems like a rather large proportion, and there is reason to believe that this value represents some sort of upper limit. The lyrics of the two verses were in this case very similar, sharing several words, which might have contributed to the confusion and increased the number of fixations on the lyrics of the first verse. Furthermore, the errors mentioned indicates that this specific subject was more dependent on the lyrics of the first verse than were the others.

The general result, however, should not merely be rejected as some sort of idiosyncrasy of a specific subject. Other subjects did make the verse confusion error, and manual inspection of the eye movement traces showed a tendency of most subjects sometimes fixating text instead of notes. Nothing was concluded about the avail of this behaviour.

[5] Previous research (e.g. Deutsch, 1970; Patel, 1998) have not settled the question if there are different buffers for music and language or not, but if there is not, music information can still not be apprehended from the text object, and little text information from the music object.

6.5 Ecological validity

The experimental situation differed from the ordinary music reading situation in some apparent respects. The choir singers are used to singing in unison with others, thus the possibility to wait for others to articulate a new note is usually there, but not so in these recordings. Crucially, any mistakes are quickly discovered and corrected in musical flow of the choir, but when singing alone it may take a longer time to notice a mistake. It will also be harder to correct such mistakes without having anyone else to listen to.

In order to compensate for the latter discrepancy, the subjects were aided by the author support singing the next few notes when confusion occurred after a mistake. Keeping up the musical flow was considered more important than the more puristic experimental approach of not intervening at all.

It is likely that nervosity may have affected the results. The experimental situation resembles an audition more than ordinary choir singing, and although no skill tests were made, the sightreading ability was obviously exposed. The recordings where the author sang along were added partly to lessen the nervosity, partly as a safety line if the sightreading skill level would have been to low for any any meaningful data to be rendered (cf. Goolsby 1994b).

The emphasis on subject comfort precluded the use of a bite bar, resulting in worse accuracy due to head movement, but hopefully better ecological validity.

The stimuli were presented on paper, and were slightly larger than usual note size. Despite the moderate enlargement of the music in this study, one of the subjects commented on the reading of the music with two verses that perhaps he was a little too close to the notes. On the other hand, the eye-movements were not nearly as sweeping during the other recordings as during the second verse of the tune with lyrics, and the statement was perhaps in part due to the strenuous task.

Previous researchers have projected the stimuli onto a screen or sometimes presented them on a computer screen (cf. Goolsby, 1989; Goolsby, 1994a; Kinsler & Carpenter, 1995; Sloboda, 1984; Rayner, 1978). The methodology of these approaches have not been throughly discussed, and some questions raises to mind. The problem with some displays is that they do not present a text with black letters on a white background, but e.g. are green (Kinsler & Carpenter, 1995). A modern computer screen can display a stimulus in a more normal fashion, but the flickering of the light may distract concentration and affect the subjects endurance.

The projection method avoids flickering light, but the room probably has to be dark in order to get a good stimulus contrast, which is a situation that few singers are used to. A worse deficit that projection might suffer from is making the stimulus significantly larger than usual. Parafoveal and peripheral vision cannot play their usual roles in such a scenario, and once again it would be an unusual sightreading situation to read notes from a large screen.

6.6 Tempo

The tempo was not given, and the subjects had to rely on their own musical experience to determine a good tempo. Goolsby (1989) remarks that early music reading research "suffered" from lack of tempo control, and he chose to provide the tempo with a metronome in his own sightreading experiments. However there were legitimate reasons not to provide a tempo. For example, musical interpretation requires the possibility of individual tempi. Furthermore, it was considered unwise to provide too much information before a performance, thus risking the subject confusion.

The tempo differences between subjects have several reasons. To start with, a tune may, of course, be sung at somewhat different tempos and still be of high musical quality. However, the differences of the combined stimulus were larger than such artistic interpretations permit. No one performed the tune with lyrics too fast, but some subjects chose a slower tempo than it should be. The other tune admits a wider range of tempo variation and no performance was notably too fast or too slow.

The most probable explanation of the variations is that a slower tempo deliberately was preferred in order to facilitate the sight-reading task. The variation of performance times was considerably less when the author sang along.

Interestingly, the correlations between the total performance times in table 3 indicate that an individual subject prefers to perform at a certain speed compared to others, regardless of the stimulus, be it text, music or a combination of both. Surprisingly, it seems that the too slow tempi were not primarily chosen to compensate for poor musical skill, but rather to fit the subject's ordinary speech tempo. The fact that variation of performance times was less when the author thus receives another explanation: the subjects adjusted their tempi to those preferred by the author, not to "correct" tempi rejected earlier for their greater difficulty.

The eye-voice temporal distance correlated with the total time of performance in music, but not in text. The correlation was significant but not strong, and it is difficult to draw any conclusions about this result. If the tendency shown by Goolsby (1994b) that skilled readers have larger eye-voice temporal delay than poor readers is correct, then the correlations indicate that poor readers choose a faster tempo, which seems puzzling.

Some subjects increased their performance times by making a lot of pauses, hesitating about the next note. Apparently, correct intervals was considered more important than correct duration of pauses. This preference might reflect that the subjects consider correct pitch being a more important trait of high quality musical performance than correct tempo. A different but equally probable assumption is that when a new song is encountered, it is beneficial for future performance to concentrate on the dimension of pitch.

6.7 Tonality errors

To the subjects who made the mix-ups between major and minor it seems that the highest priority was to stay within a scale based on the present keynote. Previous research states that mode is more important than pitch (Besson & Friederici 1998), which explains why the subjects did not notice their changes between major and minor: they would rather keep the notes within the present mode than performing the correct intervals and change the mode back. This explanation also accounts for the faily frequent tonal answer errors, where it often took the subject several bars to notice that anything was wrong, e.g. by performing a correct interval resulting in a note out of key (the regression aimed at an accidental at the beginning of a staff was also performed directly before such an interval). Others just remembered a pitch correctly, despite the interference of several notes, which indicates that the claim by Deutsch (1970) about melody coding does not have sufficient explanatory power.

In the normal prima vista singing situations of the subjects these tonality errors are most unlikely to occur, since the harmonies of other parts provides clear indication of major or minor.

6.8 Cognitive load

The confusion of lyrics indicates that subjects rely more on text expectation when a tune with lyrics is sung prima vista than they do in ordinary language use. The expectations are always important in language reading (Besson & Friederici 1998), but the peculiar pronunciation error that was made suggests that one further conclusion can be drawn: sightreading music with lyrics must be a task of high cognitive load, since it confuses the normal language use of some subjects. The error frequency was on the whole the highest for the the combined stimulus.

The mean fixation duration was as expected the longest in the case of a purely musical stimulus, which reflects a higher cognitive load (Recarte & Nunes, 2000) than for the purely textual stimulus. However, in music with lyrics were the workload ought to be even higher, the fixations are shorter than in music without lyrics. This implies that the effect that information distributed over a larger area evokes shorter fixations (ibid.) dominates in this case, especially during the second verse. In the text reading, however, the short fixation times were probably mainly due to low saccade direction workload, since it is easy to predict where the scan path should be headed in this case (in the music reading, the vertical dimension did manifest itself more, e.g. as text dips and contour tracing).

7 Future research

The relationship between the eye-voice distance, the attention-voice distance and the attention-eye distance deserves to be investigated further. A study comprising eye-tracking as well as the window paradigm and the traditional eye-voice span technique would be of particular interest to confirm the assumptions illustrated in figure 13. In particular, in all these experiments the reading should be vocal, since it would be difficult to design a study on e.g. the silent attention-voice distance. The window paradigm has yet to be applied on vocal language and music reading.

The temporal measurements of the eye-voice distance have in the present study rendered interesting results and a spatial measurement, which can be done just as easily in a video editing program, might cast additional light on the entity. Previous eye-voice span research has dealt with the attention-voice spatial distance, but if the vocalizations are recorded, there are no hindrances to temporal measurement of the attention-voice distance. Furthermore, studies where a spatial distance is compared with its temporal correlate could make conclusions about the nature of the immediate memory and the integrative buffer possible, for both language and music. The results of this study suggest that that the temporal distance is the better choice, and furthermore that music with lyrics seperated from the score offers a different window on the music and language buffers than does music with lyrics directly below each staff.

Since there is a lack of experimental data on reading text aloud this might be a fruitful area to explore. For example, a recording of both vocal and silent text reading would help to cast additional light on the share of regressions, where a more direct comparison could determine whether there are more regressions in silent reading than in vocal reading or vice versa. The term regression has to be defined more realistically than in previous studies (e.g. Goolsby, 1994).

Many of the parameters in the statistics of the present study could have been related to skill as well, and subjects of a wider range of musical abilities making a skill test in a similar study could naturally render additional results. If a test could be contrived that separates different dimensions of sightreading, e.g. interval sense, rhythm and retention of key note in a scale, this would be a valuable asset to the investigations.

In order to analyze large amount of data, the automatization of data treatment is of great importance. The filter used in this study to measure the number of regressions could easily be made more elaborate, making studies with a large amount of text to be read possible to analyze. The number of regressions may, in silent as well as vocal reading mode, be studied with respect to e.g. text complexity, letter size, reading task, content coherence or reading speed.

8 References

Banton, L. J. (1995). The role of visual and auditory feedback during the sight-reading of music. *Psychology of Music, 23,* 3–16.

Besson, M. & Friederici, A. D. (1998). Language and music: A comparative view. *Music Perception, 16*(1), 1–9.

Deubel, H., Irwin, D. E. & Schneider, W. X. (1999). The subjective direction of gaze shifts long before the saccades. In: Becker et al. (ed.) *Current Oculomotor Research*. New York: Plenum Press, 65–70.

Deutsch, D. (1970). Tones and numbers: Specificity of interference in immediate memory. *Science, 168,* 1604–1605.

Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review, 85*(4), 341-3

Goolsby, T. W. (1989). Computer applications to eye movement research in music reading. *Psychomusicology, 8*(2), 111–126.

Goolsby, T. W. (1994a). Eye movement in music reading: Effects of reading ability, notational complexity, and encounters. *Music Perception, 12*(1), 77–96.

Goolsby, T. W. (1994b). Profiles of processing: Eye movements during sightreading. *Music Perception*, *12*(1), 97–123.

Hansen, J. P. (1994). *Analyse af læsernes informationsprioritering*. Roskilde: Forskningscenter Risø.

Henderson, J. M. & Hollingworth, A. (1999). High level scene perception. *Annual Review of Psychology*, *50*, 243–271.

Hoffman, J. E. (1998). Visual attention and eye movements. In: Paschler, H. (ed.) *Attention*. London: University College London Press, 119-154.

Holsanova, J. (2001). Picture viewing and picture description: Two windows on the mind *Lund University Cognitive Studies, 83*.

Hyönä, J., Lorch, R. F. Jr. & Kaakinen, J. Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. (in press). *Journal of Educational Psychology*.

Kinsler, V. & Carpenter, R. H. S. (1995). Saccadic eye movements while reading music. *Vision Research*, *35*(10), 1447–1458.

Patel, A. D. & Peretz, I. (1997). Is music autonomous from language? A neuropsychological appraisal. In: Deliège, I. & Sloboda, J. (ed.) *Perception and cognition in music*. Guilford: Psychology Press.

Patel, A. D. (1998). Syntactic processing in language and music: Different cognitive operations, similar neural resources? *Music Perception*, *16*(1), 27–42.

Pelz, J. B., Canosa, R. & Babcock, J. (2000). *Extended tasks elicit complex eye movement patterns*, ACM SIGCHI Eye Tracking Research & Applications Symposium 2000.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin*, *85*(3), 618–660.

Recarte, M. A. & Nunes, L. M. (2000) Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology: Applied*, *6*(1), 31–43.

Reichle, E. D., Rayner, K. & Pollatsek, A. (2000). Comparing the E-Z reader model to other models of eye movement control in reading. *Cogprints*, <http://www.cogprints.soton.ac.uk> (Jan. 2002).

Sloboda, J. A. (1974). The eye-hand span — an approach to the study of sight reading. *Psychology of Music*, *2*(2), 4–10.

Sloboda, J. A. (1984). Experimental studies of music reading: A review. *Music Perception*, *2*(2), 222–236.

Sloboda, J. A. (1985). *The musical mind*. New York: Oxford University Press.

Yarbus, A. L. (1967) *Eye movements and vision*. New York: Plenum Press.

Acknowledgements

Thanks to the choir singers who participated as subjects, thereby making this paper possible. Special thanks to Kenneth Holmqvist for generously sharing his time, experience and wise advice.

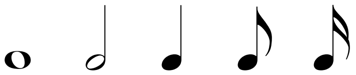


Figure 14: *Common notes. Starting from the left, we have the semibreve, the minim, the crotchet, the quaver and the semiquaver. Each of these notes have half the duration of the note to its left.*

A Appendix: Introduction to standard western musical notation

A.1 Staffs and notes

An example of written music is shown in figure 3 in page 20. The different rows are called *staffs*, each staff consisting of five lines covered with notes. The whole set of staffs constitutes a *score*. The part common to all note symbols is the oval *note head*, which vertical position determines the pitch of the note. Some notes have *stems*, vertical lines attached to the head, and at the end of the stem sometimes have a *tail*, which resembles a flag. In some cases tails of adjacent notes are joined into *bar lines*.

The duration of a is concluded from the details of the symbol. This is shown in figure 14, where the British names for the notes also are presented. Musical pauses of different duration have different appearances as well. A dot added after a note or a pause increases its time value by 50 per cent.

A.2 Clefs and intervals

The relation between two note pitches is referred to as an *interval*. This relation is not absolute frequency differences, but proportional relations, the *octave* being a fundamental interval where the frequency relation is two to one. The Western sets of notes used together, *scales*, divides the octave into twelve half tone steps, each scale utilizing eight of the notes in an octave. Thus, the steps between notes are sometimes half tones, sometimes whole tones. The intervals within an octave are named from the order of the scale notes which form each interval when compared to the first note of the scale, the *keynote*, e.g. a second and a fifth. Some intervals that share their names may still differ a half tone, and they are specified as *minor* or *major*. Those who cannot differ this way are called *perfect*.

The ranges of different instruments and voices spans over different frequency intervals, therefore the reference tone associated with a specific vertical position can be altered. The key to this is the adorning *clef* at the beginning of each system. In this case the *treble clef* is used, implying that the second line from the bottom corresponds to a *middle G* with a frequency of 415 Hz. The other lines and the spaces between them are assigned pitches accordingly, so that a rising sequence of notes, *scale*, is formed when the staff is traversed upwards.

In choir music the treble clef is used for female voices, while basses, and sometimes also tenors, make more use of the *bass clef*. However, most male singers are used to the treble clef as well[6], by convention performing the notes one octave lower (at half the frequency) than written.

A.3 Key signature

Normally, the notation corresponds to the white keys of the piano, the *major C scale*. Whenever a black key note is to be performed, this is indicated by an *accidental*, as the double crosses, *sharp signs*, in figure 3. If a musical piece is composed in another scale, the accidentals would have to be written in front of every note of a certain vertical position. In order to avoid this, the *key signature* is shown beside the clef; these accidentals should be applied to every note of the same vertical position.

A.4 Bars and time signature

The *bars* are short parts of the piece, divided by vertical *barlines*. The fraction at the beginning of a song is the time signature, where the number of beats in each bar is written above the length of one beat (taken as a fraction of a semibreve). Thus, the tune in figure 1 is notated with three beats in each bar, each beat holding the time value of a crotchet.

A.5 Additional music notation features

A note roofed by a *fermata*, i.e. half a circle with a dot in it's centre, is of optional duration. Other tempo adjustments are notated with text (most commonly Italian), as the *rit.* in figure 3, standing for *ritardando* or *gradually slowing down*. Dynamical changes are in the music at hand only denoted by the letters *p* and *f*, standing

[6] E.g. church psalms and other songs in unison are notated in treble clef.

for *forte* (loud) and *piano* (soft), respectively. Mood or suggested tempo may be notated directly above the first bar of the piece.

B Appendix: Music and language perception

Sloboda (1985) points out that both music and language are features universal to all humans and specific to humans. Chimpanzees cannot play instruments, even though they would love to try, and the musical abilities of humans are just as fascinating as language. The language is extensively used as a window of the mind, and perhaps music has just as much to say.

A brief survey of differences and similarities between music and language is presented below, the objective of which is to motivate the comparative approach of this study.

B.1 Modularity discussion

Anyone who is concerned with the relationship between language and music has to consider two basic questions: Firstly, are there features of language which does not exist in music, and, secondly, are there features in music which cannot be accounted for in language studies?

To start with the first question, it would be foolish, according to Sloboda (1985) to say that music is another natural language. Clearly there are some differences, e.g. we use language to exchange information about the world and the relations in it with each other. On a syntactic, structural level, grammatical categories in music, (such as chords and intervals), have little in common with those of language (such as noun and verb phrases) (Patel, 1998).

The second question is of particular interest for experimental music research, because if there is no essential differences between the two modes of expression, there would be no necessity of breaking new ground in the complex field of music, when so much research already has been conducted on language.

One argument takes as a starting point the tendency that language is more conventionalized than music. Even though musicians seldom disagree on local features of a piece (such as tone duration and pitch), they often have different opinions on the global features (such as phrases), and the opinion of the same person is not constant over different encounters with the music. Raffman (1993) argues that since the knowledge displays itself in different ways every time, it cannot be expressed and understood by means of a language. Thus she claims that, unlike language, music is in some respects *ineffable*, i.e. you cannot express all musical knowledge in words. Sometimes a musician “just knows” that a certain phrase should be performed a little slower, without apparent explanation. As a matter of course, most musical features are *effable*: the *fa* note calls for a subsequent *mi* note in a C major piece, the tempo should go down as one approaches the end of a piece, etc. The consistency of local feature reports implies, according to Raffman (ibid.), consistency in the mental representations of music, and the inconsistency of global reports implies likewise that no durable representations exist on this level.

B.2 Structural similarities

On a basic level, similarities between language and music are obvious. Fundamental units — the phonemes and the notes, respectively — are characterized by frequency and duration parameters. Combinations of these units offer an infinite number of possible utterances, in speech as well as in music performance.

Similarities exists in several aspects of music and language. Sloboda (1985) compares Shenker's structural theories of music with Chomsky's language theories. For example, a fundamental differentiation between surface structure and deep structure signifies both music and language.

The surface structure is the form, that may be altered without affecting the deep structure, which in the case of language would be the meaning of an utterance. Thus, different word order may comprize the same meaning. With music, the deep structure is not as easily described, but two musical phrases may have a strong resemblance even though the melody line differs slightly.

B.3 Methodological similarities

In a comparison between language and music, Besson and Friederici (1998) concludes that speech and music essentially share the same physical properties. Similar methods can be used in the study of both, which makes comparisons possible and provides an opportunity to determine which concepts are specific to one domain and which may explain both processes. They suggest that an investigation of the relationship between acoustic and prosodic properties in speech and acoustic melodic contour in music would be a fruitful approach to the research comparing language and music.

According to Besson and Friederici (ibid.), prosody has been a somewhat neglected area in linguistic studies, despite it's importance for structural information in most languages and it's lexical function in tonal languages, such as Chinese.

B.4 Recoding and meaning

Deutsch (1970) proposes, as is mentioned in section 1.3.11, that auditorial musical material is recoded in the memory storage process. This recoding process also exists in music *reading*, according to Sloboda (1984) and Goolsby (1994b). They emphasize on music reading being a case of music perception, i.e. there is no necessity to vocalize a prima vista reading in order to percieve and process the music.

Recoding in language reading is more obvious, and language perception may affect e.g. pictures in our memory. Sloboda (1985) adresses the tendency that of musical memory appears more isolated, and even though we can describe music via metaphor, the clear recoding of language information to semantic information is not as obviously applied to musical information. The case of musical meaning has popularly been under heavy debate, but is not yet settled.

Besson and Friederici (1998) points out that words carry meaning by convention, while music is more self-referential, i.e. tones having meaning mainly by reference to the preceeding ones.

Besson and Friederici (ibid.) also remarks that the common notion of music being “the language of emotions” is hardly true. They state that emotions can be expressed to a wide extent in our ordinary language. It seems that people are hoping for an easy solution to a severe problem if they say that music is the language of emotion in an attempt to find some counterpart in music for the lexical meaning in speech. There is emotion in speech as well, but also, music cannot be called a language since it holds dimensions that does not exist in any language.

References

Besson, M. & Friederici, A. D. (1998). Language and music: A comparative view. *Music Perception*, *16*(1), 1–9.

Deutsch, D. (1970). Tones and numbers: Specificity of interference in immediate memory. *Science*, *168*, 1604–1605.

Goolsby, T. W. (1994b). Profiles of processing: Eye movements during sightreading. *Music Perception*, *12*(1), 97–123.

Patel, A. D. (1998). Syntactic processing in language and music: Different cognitive operations, similar neural resources?. *Music Perception*, *16*(1), 27–42.

Raffman, D. (1993). *Language, music and mind*. Cambridge, Massachusetts: MIT Press.

Sloboda, J. A. (1984). Experimental studies of music reading: A review. *Music Perception*, *2*(2), 222–236.

Sloboda, J. A. (1985). *The musical mind*. New York: Oxford University Press.

C Appendix: Translations

C.1 The title of the purely musical stimulus

The old man was about to take the cows to pasture

C.2 The lyrics of the music and language stimulus

The he-goat that you gave me

1. The he-goat that you gave me, I have it.
If I can have another, I take.
Goats and lambs I give to you.
Say, say, say, will you have me?
Have yourself a goat!

2. The goat that I gave you, you have it.
If you can have another, you take.
The goat, well, I give it to you.
Say that you got it from me.
Have yourself a he-goat.

C.3 The lyrics of the purely textual stimulus

Now the lovely song has gone silent

1. Now the lovely song has gone silent,
that I heard so often amongst the voices of the birds.
Now I miss the lovely words,
that you spoke and said as comfort for my heart.
No happiness for my heart,
no relief for my distress,
for I see that your love is at an end.

2. You have made your love disappear,
like the clouds they disappear upon the sky of blue.
It has made my heart feel,
that it was your full intention that you would reject me.
Because so many charming boys,

and so sweet and they speak,
and I know that they betray me still.

3. Try to get yourself a friend amongst the rich,
they have money to palliate a thousand flaws.
There you will for certain find your equal,
among the poor girls you will not be concerned.
But then all riches,
I wish to compare to a flower,
that by even withers away.