Voice and Stream: Perceptual and Computational Modeling of Voice Separation

Emilios Cambouropoulos
Aristotle University of Thessaloniki, Greece

listeners are thought to be capable of perceiving multiple voices in music. This paper presents different views of what 'voice' means and how the problem of voice separation can be systematically described, with a view to understanding the problem better and developing a systematic description of the cognitive task of segregating voices in music. Well-established perceptual principles of auditory streaming are examined and then tailored to the more specific problem of voice separation in timbrally undifferentiated music. Adopting a perceptual view of musical voice, a computational prototype is developed that splits a musical score (symbolic musical data) into different voices. A single 'voice' may consist of one or more synchronous notes that are perceived as belonging to the same auditory stream. The proposed model is tested against a small dataset that acts as ground truth. The results support the theoretical viewpoint adopted in the paper.

Received October 31, 2007, accepted May 13, 2008.

Key words: voice separation, voice leading, stream segregation, auditory streaming, computational modeling

The notions of 'voice,' as well as, homophony and polyphony, are thought to be well understood by musicians, and listeners are thought to be capable of perceiving multiple voices in music. However, there exists no systematic theory that describes how voices can be identified, especially when polyphonic and homophonic elements are mixed together. This paper presents different views of what 'voice' means and how the problem of voice separation can be systematically described, with a view to highlighting various aspects of the problem and developing better cognitive and computational models of music. Vague (or even contradicting) treatments of this issue will be presented.

Voice separation refers to the task of separating a musical work consisting of multi-note sonorities into independent constituent voices. Recently, there have been a number of attempts (e.g., Cambouropoulos, 2000; Chew & Wu, 2004; Kilian & Hoos, 2002; Kirlin & Utgoff 2005; Madsen & Widmer, 2006; Szeto & Wong, 2003; Temperley, 2001) to model computationally the segregation of music into separate voices. Much of this research is influenced by empirical studies in music perception (e.g., Bregman, 1990; Huron, 2001; McAdams & Bregman, 1979), as well as by more traditional musicological concepts such as melody, counterpoint, voice-leading and so on. It appears that the term 'voice' has different meanings for different research fields (traditional musicology, music cognition, computational musicology); especially in the computational domain, researchers sometimes adopt oversimplified views on this topic that lead to limited results.

Voice separation algorithms are very useful in computational implementations as they allow preprocessing of musical data, thus opening the way for more efficient and higher quality analytic results. In domains such as music information retrieval or automated musical analysis, having sophisticated models that can identify multiple voices and/or musical streams can assist more sophisticated processing within the voices (rather than across voices). For instance, if one wants to identify musical works that contain a certain melodic pattern, this pattern should not be spread across different parts (perceptually implausible) nor in voices that are not perceptually independent (e.g., internal parts in a homophonic work), but within voices that are heard as having a life of their own.

In this paper, we first will discuss multiple meanings of the term 'voice' that often create confusion when one attempts to model voice separation. Second, the notion of voice is examined more specifically in relation to the notion of musical stream primarily in the context of David Huron's recent study of voice-leading and auditory streaming. Then, cognitive principles are outlined that are considered to be essential for voice/stream segregation. Finally, existing computational models for voice separation are reviewed, and a new algorithm is presented.

Music Perception volume 26, issue 1, pp. 75–94, issn 0730-7829, electronic issn 1533-8312 DOI:10.1525/mp.2008.26.1.75

What is a Voice?

Before looking into various aspects of voice separation, it is important to discuss what is meant by the term 'voice.' A few musical examples will assist our inquiry.

In Figure 1a, a short passage from Bach's Chaconne for solo violin (mm. 33-36) from the D Minor Partita (BWV 1004) is depicted. This passage is considered a monophonic passage, i.e., a single voice, since it is performed by a solo violin. At the same time, this is a case of 'implied polyphony' or 'pseudopolyphony,' where the lower descending chromatic sequence of tones may be separated from the higher tones leading to the perception of two independent voices (Figure 1b). Finally, this succession of tones can even be separated into three different voices (Figure 1c) if the implied triadic harmony is taken into account (Figure 1d).

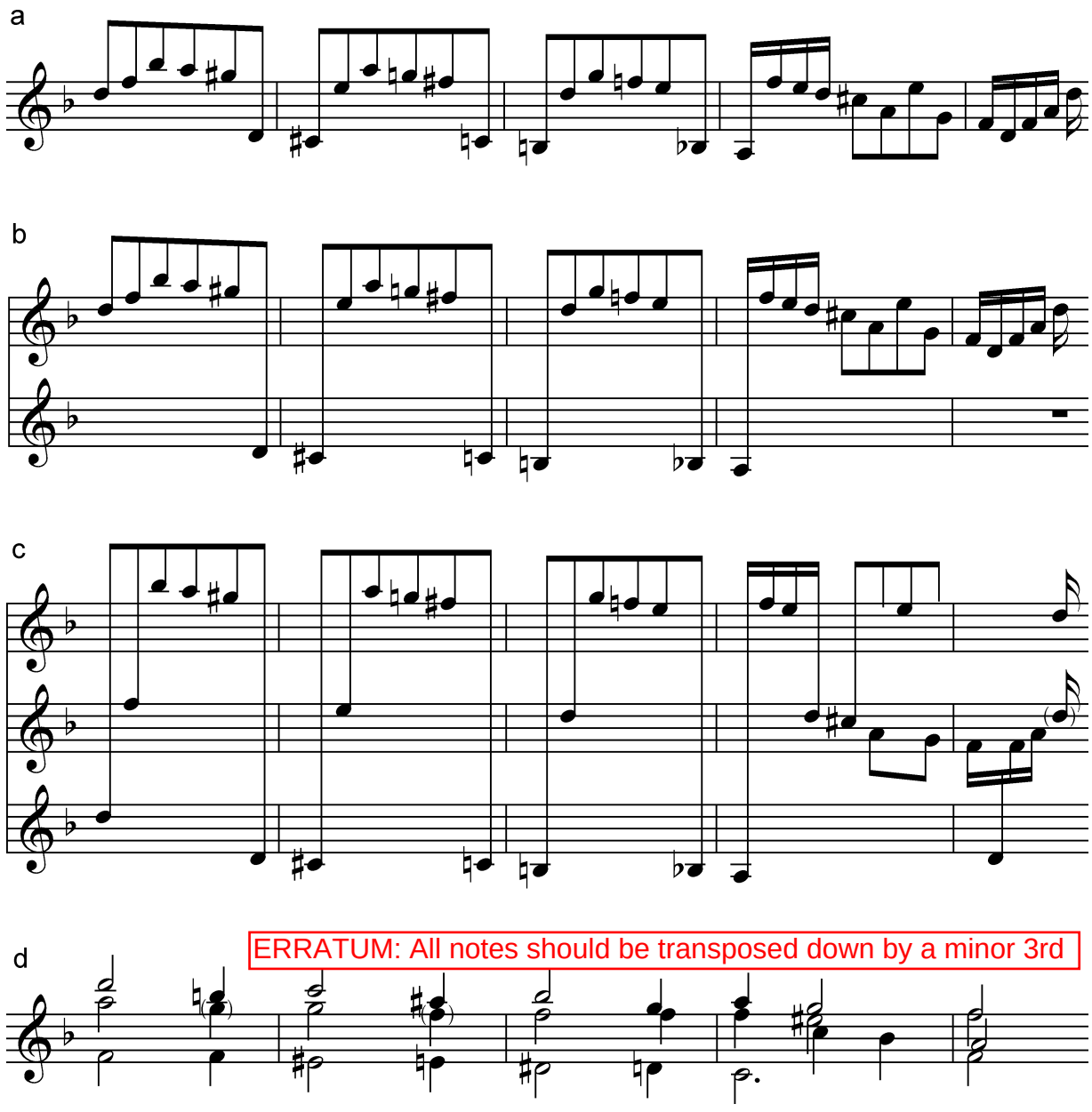

FIGURE 1. Measures 33-36 from Bach's Chaconne for solo violin from the D Minor Partita (BWV 1004) presented as: (a) one voice (solo violin); (b) two voices (perceived implied polyphony); or (c) three voices following the implied triadic harmonic structure (harmonic reduction presented in (1d).

For this musical passage, we can see three different ways in which 'voice' may be understood: (a) literally instrumental 'voice,' (b) perceptual 'voice' relating to auditory streaming, and (c) harmonic 'voice' relating to harmonic content and evolution. In practice, there is often significant overlap between these meanings; however, there are many cases in which these notions are incongruent. Before examining the relations between these different meanings, each of these will be briefly discussed.

1. In compound words/terms such as monophony, polyphony, homophony, and heterophony, the second constituent part ('-phony') comes from the Greek word 'phōnē' (φωνη´), which means 'voice' (of humans or animals) or even 'the sound of musical instruments.' [1] In this sense, the term voice is used primarily to refer to the sound sequences produced by different musical sound sources such as individual choral voices or instrumental parts (e.g., in string quartets, wind quintets, and so on). In this sense, the passage of Figure 1 is literally monophonic since it is performed by a solo violin.
2. Auditory stream integration/segregation (in music) determines how successions of musical events are perceived as belonging to coherent sequences and, at the same time, segregated from other independent musical sequences. A number of general perceptual principles govern the way musical events are grouped together in musical streams. Bach's monophonic passage can be perceived as consisting of two independent musical sequences/streams (Figure 1b); the lower tones may be integrated in a descending chromatic sequence of tones primarily due to pitch proximity—however, as the tempo of this passage is rather slow, segregation is usually enhanced via dynamic and timbral differentiation of the two sequences during performance (see extended study on implied polyphony by Davis, 2001).
3. Since a monophonic passage commonly implies a specific harmonic structure, the tones that correspond to the implied chords may be considered as implying a horizontal organization into a number of separate voices. Dann (1968) separates a brief passage from Bach's B minor partita into multiple voices (up to five voices) based on melodic, harmonic, and rhythmic aspects of each tone. He does argue, however, that the five voices need not be perceived. In the current example, Bach's monophonic passage implies essentially a triadic harmonic structure[2] such as the one presented in Figure 1d. If such a harmony is perceivable, one could hypothesize that a listener is capable of organizing tones vertically in chords and that the tones of these chords imply multiple horizontal lines or voices: in this case, three voices as shown in Figure 1c (the end of the passage may be separated into four voices—downward stems in the third stave indicate a fourth voice—this is one possible harmonic voice separation among others).

The first literal meaning of the term 'voice' may be broadened, making it applicable to music produced by a single 'polyphonic' instrument (such as the piano, celesta, guitar, etc.)[3] in cases where the music consists of a relatively fixed number of individual musical lines (e.g., 3- or 4-part fugues or other 4-part works for keyboard instruments). In such cases, the music can be thought of as comprising a number of concurrent monodic lines or 'virtual' voices. The horizontal motion of individual voices from note to note in successive chords is governed by the rules of voice-leading (at least for a large part of Western art-music). Terms that are synonymous to 'voice' in this sense are 'part' and 'line'—in the case of polyphonic music the term 'contrapuntal voice' is often used.

This meaning has limitations in regards to perception as it is possible to have monodic musical lines splitting into more that one perceptual stream (e.g., in the case of 'implied polyphony'), or, conversely, different individual voices merging into a single stream (e.g., homophonic accompaniment). The musicological meaning of voice is not fully congruent with a perceptually oriented meaning of voice (relating to auditory streaming). Perceptual factors are sometimes taken into account implicitly by music theorists, but the distinction between the two notions, i.e., 'voice' and 'stream,' is not always clear. Implied polyphony is a relatively rare case where musicologists/music theorists explicitly resort to music perception.[4] Or melodic lines moving in parallel octaves are commonly considered (in acoustic and perceptual

[1]Liddel and Scott, *Greek-English Lexicon.* Oxford University Press.

[2]Harmonic analysis provided by music theorist/analyst Costas Tsougras.

[3]This applies more generally, not only to polyphonic instruments, such as piano and guitar, but to groups of instruments that produce timbrally undifferentiated polyphonic textures, such as string quartets, choral groups, brass ensembles, and so on.

[4]For instance, Swain (2002, Chapter 6) discusses various musicological factors that determine densities of harmonic rhythm (relating directly to number of voices). The only time he refers explicitly to perception is when dealing with implied polyphony (compound melody): "Since the present criterion for density is the number of voices, each *perceived* voice counts, though they do not arrive precisely coincident with the new triad" (p. 65).

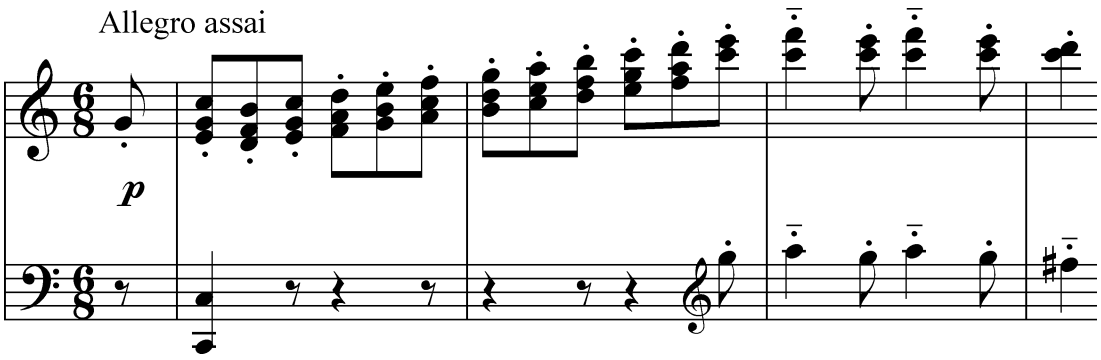

FIGURE 2. Opening measures of Beethoven's Sonata Op. 2, No. 3. Should this passage be understood as determining three parallel voices or a single chordal voice?

terms) as a single 'amplified' voice. There is need for a greater clarity on how voice separation relates to stream segregation (see next section).

Implicit in all of the descriptions in relation to the above example is the assumption that 'voice' is essentially a *mono*phonic sequence of successive nonoverlapping musical tones. Traditional voice-leading involves rules that govern the note-to-note movement between chords. These note-to-note links determine individual voices that are monophonic since they contain sequences of single tones. In general, a single voice is thought not to contain multiple-note sonorities.

Let us consider the opening few measures of Beethoven's Sonata Op. 2, No. 3 (Figure 2). In this passage one sees three voices moving in parallel. In perceptual terms, however, one could argue that a listener hears essentially one stream of ascending parallel $^{6}_{3}$ chords (in a sense, one 'thickened' voice[5]). It is clear that the perception of a single musical stream is prior to the perception of each of the individual parallel lines of tones. It could even be argued that it is actually hardly possible to perceive independently the individual constituent lines at all (especially the 'inner' middle voice or even the lower voice). The perceptual principles that merge all these tones into a single auditory stream will be discussed in more detail in the following sections. It is important, however, to note at this point that theoretical and perceptual aspects of voice may occasionally contradict each other and that the monophonic definition of voice may require rethinking.

[5]"Two voices in parallel motion are melodically less independent than otherwise, and may be looked upon as a single voice with duplication" (Piston, 1991, p. 29).

The excerpt from Rachmaninov's Prelude Op. 3, No. 2 (Figure 3) presents an example where both fission and fusion of voices appear concurrently (i.e., both pseudopolyphony and homophonic merging). As this passage is performed at a very fast tempo, it is perceived as two musical streams, i.e., an upper stream of descending 3-note chords and a lower stream of 2-note chords (repeating pattern of three chords). In terms of voices, it could be argued that the passage consists of five voices split into two homophonic strands. It is clear, however, that a listener does not perceive five independent voices but rather organizes the notes into two streams (as indicated by the cross-staff notation given by the composer). This example illustrates possible incongruence between the music theoretic notion of voice and the perceptually based notion of auditory stream.

In the context of this paper, the term 'voice' (usually in single quotes) will be taken to refer to *perceptually independent* sequences of notes or multi-note simultaneities. In this sense, the musical extracts in Figures 2 and 3 could be described as comprising of one and two 'voices,' respectively. Terms such as musical line or musical stream are also used to refer to such sequences. Such terms may be more appropriate in order to avoid confusion with more traditional musicological uses of the term 'voice.' In this text, however, we will use the term 'voice' (along with the other terms) in an effort to provoke further discussion and, potentially, to clarify and broaden the meaning of the musicological term.

'Perceiving' independent sequences/streams can mean different things. It could mean to be able to focus and follow a certain stream (i.e., to bring a sequence of tones to the perceptual foreground) or to follow more than one concurrent stream together, or to be able to say how many streams co-exist at a certain point (for instance, Huron, 1989, has studied voice denumerability

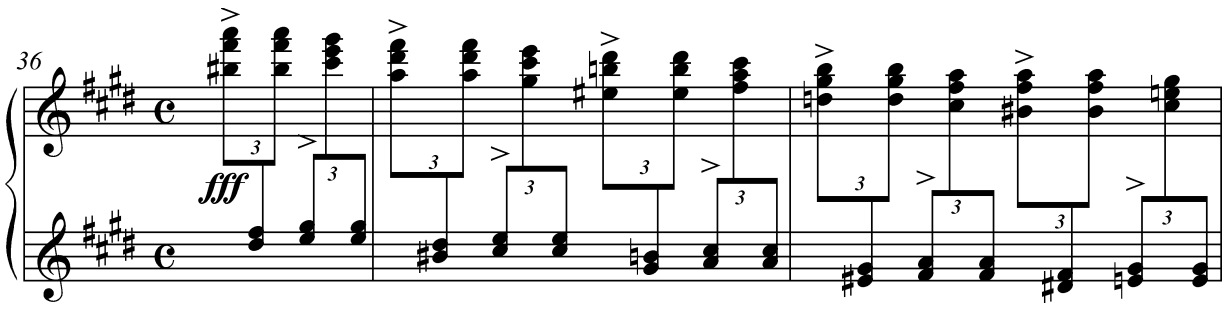

FIGURE 3. Excerpt from Rachmaninov's Prelude Op. 3, No. 2 (mm. 36-38).

in polyphonic music). Additionally, there is a question regarding the kind of listener assumed, i.e., is it an 'ideal' listener with extended music training/expertise, or an 'average' listener that uses primarily general cognitive capacities to organize musical material? It is assumed that there is no 'correct' way of perceiving musical structures: there exist many different levels and modes of perceiving. In the context of this paper, however, we take 'perceiving' to mean the ability of an 'average' listener to bring to the foreground a coherent sequence of sonorities (individual tones or combinations of tones) in a relatively effortless manner (without special training and expertise). Many of the assertions made in the paper are based on general musical knowledge and intuition (rather than concrete emprical research—further empirical research is necessary) with a view to developing computational tools that can be useful for various tasks.

Despite the importance of timbral and spatial characteristics of sound in stream segregation (Bregman, 1990), we will not discuss these aspects in the current study, as we will focus on stream segregation in timbrally and spatially undifferentiated music. An interesting aspect of music in terms of auditory scene analysis is that sound emitted from a single musical source (e.g., piano) can be organized perceptually into multiple auditory streams. This paper will focus on this aspect of musical stream segregation.

In the next section, the relationship between the notion of voice and stream is discussed in more detail.

Voice and Stream

David Huron's seminal paper 'Tone and voice: a derivation of the rules of voice-leading from perceptual principles' (Huron, 2001) is aimed at explaining "voice-leading practice by using perceptual principles, predominantly principles associated with the theory of auditory stream segregation" (p. 2). The paper develops a detailed exposition in support of the view that "the principle purpose of voice-leading is to create perceptually independent musical lines" (p. 2) and links "voice-leading practice to perceptual research concerning the formation of auditory images and independent auditory streams" (p. 6). Huron presents a set of 10 perceptual principles (6 primary and 4 auxiliary) and shows how these (essentially the 6 primary principles) may explain a large number of well established voice-leading rules. The paper gives a broad survey of empirical research primarily from the domain of auditory streaming and also presents numerous statistical analyses of actual musical data that are in agreement with the perceptual principles.

Recent computational research in voice separation modeling refers to this paper as it provides an authoritative survey of relevant auditory streaming principles, and explicitly links voice and voice-leading with perceptual principles. Researchers take as a starting point one or more auditory streaming principles and then develop computational models that are tested on a set of multivoice musical works.

The discussion below examines the relation between voice and stream from a somewhat different viewpoint than Huron: rather than taking traditional voice-leading rules as a given and explaining them through perceptual principles (mostly auditory streaming principles), the current study starts with auditory streaming principles and re-examines the notion of voice per se. Seen from this perspective, some of Huron's claims and assertions are reinterpreted, and the underlying auditory principles are organized in a somewhat different way, so as to reveal a closer link between the notion of auditory stream and a broadened perceptually based meaning of voice.

Despite the fact that Huron's paper discusses in detail voice-leading and auditory streaming processes, there is no clear definition of what voice is and how it relates to an auditory stream. Implicit in the whole discussion

seems to be that voice is a monophonic sequence of notes (this comes directly from the description of voice-leading as a set of rules that pertain to the horizontal movement from "tone to tone in successive sonorities," p. 2), and that a voice is a kind of auditory stream (i.e., a voice is a special case of the broader concept of an auditory stream that pertains more generally to all kinds of musical and nonmusical auditory events).

Perhaps the most important question is whether 'voice' is always a perceptually pertinent notion, or whether it is a music theoretical notion that in certain circumstances (but not always) may be perceived as an independent sequence of tones. If it is the former, then the link with auditory streaming occurs rather 'naturally' but the question is shifted towards determining perceptually distinguishable sequences of notes. In this case, for instance, one should not talk of 'voices' in homophonic music as it is implausible that a listener follows individual inner parts in timbrally homogeneous homophonic music (a listener tends to hear primarily a melodic line and an accompanying harmonic progression—in many occasions it is hardly possible to follow an inner part and, if it is, it requires special attention/effort on the part of the listener). If it is the latter, the link between voice and auditory streaming is less direct and the explanatory power of the perceptual principles is diminished. This means that the perceptual principles may explain why voice-leading rules came about in the case of polyphonic music (where voice independence is strong) but these principles need not always be in agreement with voice-leading rules (voice-leading rules do not always determine independent sequences of tones, as in the case of homophonic music).

It is not clear which of the above two views Huron endorses. Huron believes that "the principal purpose of voice-leading is to create perceptually independent musical lines" (p. 2). This does not imply, however, that voice-leading *always* achieves this purpose (for instance, despite compliance with voice-leading rules, musical lines are not truly independent in homophony). In this sense, he may be closer to the second view that voice and stream may be partially incongruent. On the other hand, the fact that no fundamental distinction is made between voice-leading in polyphony and homophony (these are considered specific 'genres' that are optional and appear as a result of auxiliary principles) seems to imply that voices are always perceptually independent if only to a lesser degree in homophonic music.

In the light of the onset synchrony principle (i.e., perceptual independence of parts is assisted by onset asynchrony between their notes) and the assumption that "homophonic voice-leading is motivated (at least in part) by the goal of stream segregation—that is, the creation of perceptually independent voices," Huron wonders "why would homophonic music not also follow this principle? . . . why isn't all multipart music polyphonic in texture?" (p. 44). He then proposes two additional perceptual goals in the case of homophony (namely, preservation of the intelligibility of text and/or rhythmic uniformity associated with marches, dances, etc.) that have priority over stream segregation and may account for "this apparent anomaly" (p. 44).

Rather than regarding the use of onset synchrony in homophony as an 'anomaly,' it may make more sense to question, in the first place, the assumption that voice-leading aims at creating perceptually independent voices. Under the entry 'part-writing' (which is the British equivalent of 'voice-leading') in the New Grove Dictionary of Music and Musicians, Drabkin (2007) suggests that voice-leading is "an aspect of counterpoint and polyphony that recognizes each part as an individual line, not merely an element of resultant harmony; each line must therefore have a melodic shape as well as a rhythmic life of its own" (p. 258). If rhythmic independence is considered an integral part of voice-leading, then it makes sense to assume that voice-leading aims at creating independent voices. If, however, voice-leading relates solely to pitch-to-pitch movement between successive chords (as is commonly accepted by music theorists and endorsed by Huron), then its principal purpose cannot be merely "to create perceptually independent musical lines" (Huron, 2001, p. 2). Independency of melodic lines is supported both by rhythmic and melodic factors. If one insists in giving primacy to one of these two factors, it is more plausible that rhythmic independence actually is the principal parameter—this is supported by the fact that rhythmic differentiation between voices is probably the most important discriminating factor between homophony and polyphony.

The question arises whether traditional voice-leading as a whole (seen as note-to-note movement in successive sonorities) contributes to voice independence, whether a certain subset of voice-leading rules plays a primary role or, even, whether some nontraditional rules are significant in voice segregation. It is true that traditional voice-leading rules contribute to giving parts an individual melodic shape, but it is herein suggested that melodic shape alone is not sufficient for voice independence. Consider, for instance, the two musical examples in Figures 4 and 5. In the first example (Figure 4), we have homophonic writing by J. S. Bach which, despite the relative 'independence' of the four voices, is readily perceived as a single auditory stream that consists of a melody and accompanying

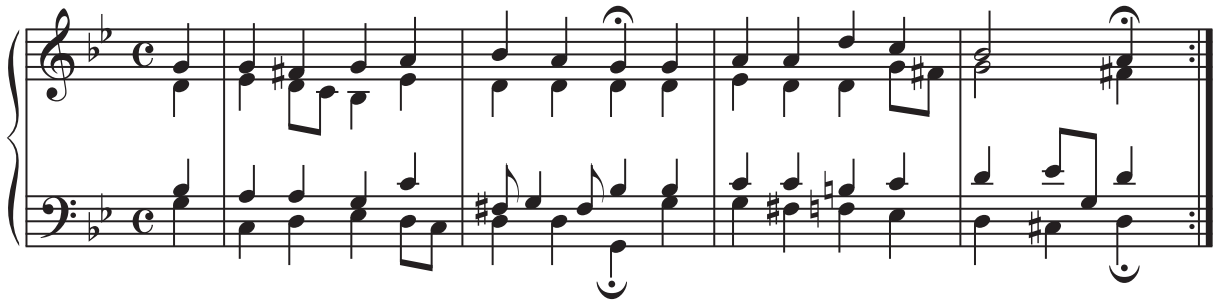

FIGURE 4. Opening of Chorale 73 *(Herr Jesu Christ, du höchstes Gut)* by J. S. Bach.

harmony—inner voices are very difficult to follow and even the bass line is not meant to be heard in the foreground. In this case, traditional voice-leading results in a perceivable musical *texture*, not independent musical lines. What matters is not individual 'threads' of tones but, rather, the overall 'texture' of the homophonic musical 'fabric.' In the second example by Beethoven (Figure 5), a listener clearly hears two streams moving in opposite directions in the second parts of the first three measures, as well as in measure 17, even though these segments are of a 'homophonic' nature (synchronous onsets among the notes of the three or more 'voices'). In this case, voice-leading is rather 'nontraditional' as parallel or similar movement between voices for relatively long periods is usually avoided. See, also, the example of two streams of block chords in Figure 6; this is a case of homophony where common-practice considerations of voice-leading are disregarded.

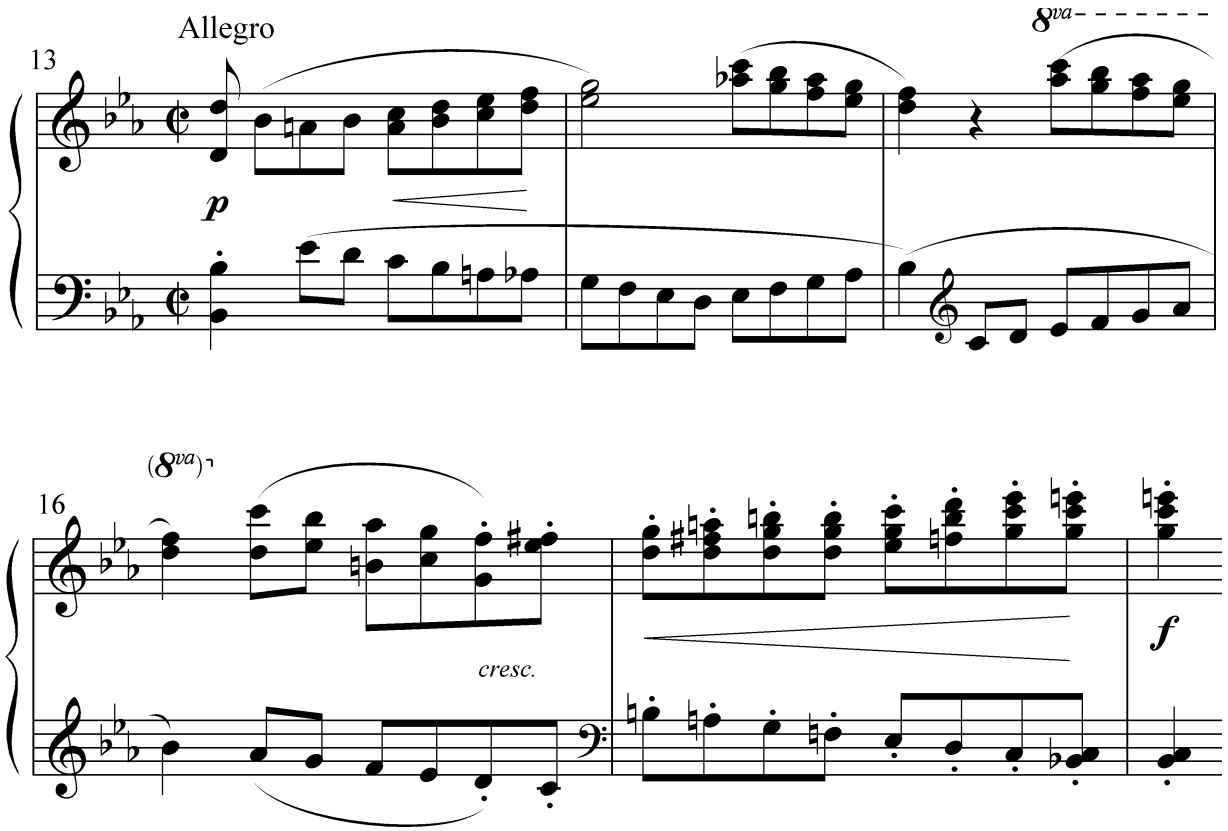

FIGURE 5. Excerpt from Beethoven's Sonata Op. 81 *Les Adieu* (mm. 13-18).

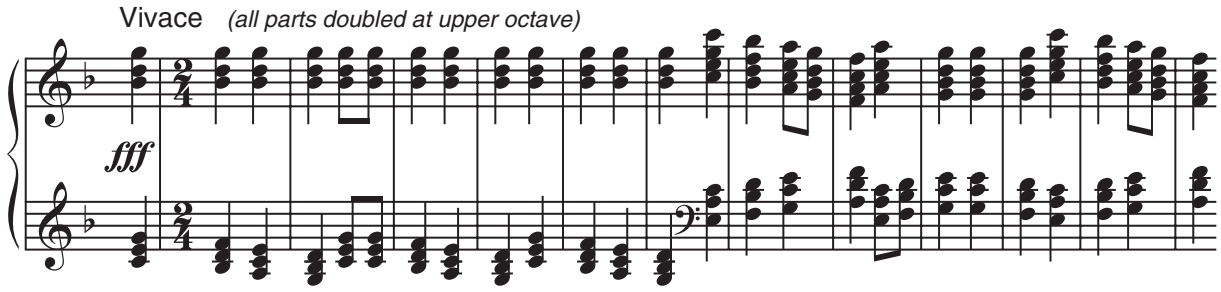

FIGURE 6. Two streams of 'block chords' in Stravinsky's, Petrushka, Tableau I. This excerpt can be regarded as a combination of two lines thickened by block chords (Piston, 1991, p. 488).

Seen from a different viewpoint, one could actually argue that traditional voice-leading in homophony 'guarantees' that no voice (apart from the upper-part melody and possibly the bass line) may be perceived independently within the overall texture. In a sense, voice-leading 'survives' in traditional homophonic part writing (rhythmic independence is abandoned), but the goal now is to construct a specific homogeneous musical texture that is more than the sum of its individual parts. If traditional voice-leading rules are not followed, it is simple to construct 'pure' homophonic structures within which a listener can discern independent streams (as in the examples presented in Figures 5 and 6). Compliance with traditional voice-leading rules, however, ensures that independent parts are woven together in such a manner that overall harmonic texture emerges prior to any individual musical fiber itself (except the melody).

The core hypothesis underlying David Huron's study can be summarized as follows: Voice is directly related to the notion of auditory stream. Since voice-leading rules aim at creating perceptually independent musical voices, these rules must be compatible with, or actually derivable from, perceptual principles (primarily auditory streaming principles). A set of six core perceptual principles are considered to be pertinent to understanding voice-leading; from these six principles, most of the established traditional voice-leading rules can be derived. An additional set of four auxiliary perceptual principles can be used optionally to shape music in perceptually distinctive ways, giving rise to different musical genres (e.g., homophony vs. polyphony).

A number of caveats are presented in relation to the above hypothesis. First, Huron accepts a priori that voice-leading rules aim at creating perceptually independent voices, and then selects a number of appropriate perceptual principles from which these rules can be derived; some well established perceptual principles for auditory streaming are given a secondary role and named 'auxiliary' since they are not considered central to explaining the traditional voice-leading rules. It is suggested that, in epistemological terms, it would be more valid to accept *all* relevant auditory streaming principles as empirical axioms and then show to what extent and in which occasions voice-leading rules create perceptually independent voices. Rather than selecting perceptual principles (by means of choosing only the ones that are considered appropriate) to 'fit' the preaccepted validity of traditional voice-leading rules in regards to perceptual voice independence, it would be more appropriate to accept in advance the validity of auditory streaming principles and then examine in which cases voice-leading rules lead to perceptually independent voices and in which not (for instance, traditional voice-leading rules without the 'auxiliary' principle of Onset Synchrony do not necessarily lead to perceptually independent voices).

Second, the approach taken by Huron can be interpreted as indirectly giving traditional Western art-music voice-leading rules a kind of 'natural law' status in the sense that these rules are 'compulsory' and based on 'primary' perceptual principles as opposed to 'auxiliary' perceptual principles that are optional and can be used to shape various particular musical genres or idioms. It is suggested that such an implied view is unwarranted. Some traditional voice-leading rules may correctly be thought of as being essentially universal (e.g., parallel octaves normally fuse into a single musical line); however, some other rules may be partially attributed to perceptual principles and partially to musical taste and convention (e.g., parallel fifths[6] are not necessarily

[6]In his textbook on counterpoint, Schoenberg states that, "Parallel octaves destroy the independence of parts. Against parallel fifths only tradition speaks. There is no physical or aesthetic reason for this interdict" (Schoenberg, 1963, p. 10).

fused into a single line more than parallel sixths or thirds; in some sense the musical effect they produce is rather characteristic and, therefore, avoided or accepted as a matter of taste in different musics of various places and times). It is too strong a hypothesis to assume in advance that traditional voice-leading rules have a clear perceptual aim rather than aesthetic or other culture-specific preferences.

In the current paper, perceptual principles regarding stream segregation are taken as empirical axioms that can be used to understand and describe the notion of musical voice. The intention is to understand the perceptual mechanisms that enable a listener to break music down into horizontal strands of musical events. Such strands are often coincident with the standard notion of voice. In some cases, however, the notion of 'voice' can be altered or extended so as to be congruent with perceptual concerns of stream segregation.

The aim of the paper is to explain musical stream segregation/integration with a view to developing a formal system (implementable on a computer) that is capable of achieving perceptually meaningful 'voice separation' (not necessarily to discover 'voices' indicated in the score by the composer). In the context of the computational model proposed below we will use the term 'voice' to mean a perceptually independent stream consisting of single and/or multi-note sonorities. As will be suggested in the next section, it may be useful to reorder and restate some of the perceptual principles presented by Huron (2001), if the aim is to describe systematically musical voice/stream segregation.

A voice, seen essentially as a monophonic sequence of tones, is not always perceived as a musical stream, and conversely, a musical stream is not always a monophonic voice. If, however, the notion of voice is broadened so as to mean a succession of musical events (notes or multi-note sonorities) perceived as an independent sequence, voice and stream become strongly linked together. Neither of the two approaches is 'correct'—it is merely a matter of definition. The aim of the above discussion is to raise awareness to problems of relating stream and voice, and to prompt further research that may clarify the usage and meaning of such terms.

Perceptual Principles for 'Voice' Separation

In this section, fundamental principles of perceptual organization of musical sounds into streams will be examined with a view to establishing a framework that can form a basis for the systematic description of 'voice' separation processes (which can lead to the development of computational models for such processes). As Huron's (2001) paper provides an excellent survey of relevant research, and as it presents a set of principles that cover all major aspects of stream integration/segregation, we will use a number of these principles as the starting point of our exploration. The main principles[7] that we will refer to in this and the following sections are presented succinctly below (it is assumed, however, that the reader is acquainted with these principles):

Principle of Temporal Continuity: "Continuous or recurring rather than brief or intermittent sound sources" evoke strong auditory streams (Huron, 2001, p. 12).

Principle of Tonal Fusion: "The perceptual independence of concurrent tones is weakened when they are separated by intervals (in decreasing order: unisons, octaves, perfect fifths . . .) that promote tonal fusion." (p. 19)

Pitch Proximity Principle: "The coherence of an auditory stream is maintained by close pitch proximity in successive tones within the stream." (p. 24)

Pitch Co-modulation Principle: "The perceptual union of concurrent tones is encouraged when pitch motions are positively correlated." (p. 31)

Onset Synchrony Principle: "If a composer intends to write music in which the parts have a high degree of independence, then synchronous note onsets ought to be avoided. Onsets of nominally distinct sounds should be separated by 100ms or more." (p. 40)

No distinction is made between these principles in terms of being primary or auxiliary. They are all fundamental perceptual principles that enable a listener to 'break down' the flow of musical tones into independent voices/streams. These 'voices' are not necessarily monophonic but may contain sequences of multi-tone sonorities. 'Voice' separation is not seen from a compositional viewpoint (i.e., how a composer constructs a musical piece and what rules he/she uses) but from a perceptual viewpoint (i.e., how an average listener organizes musical tones into coherent auditory streams or 'voices').

[7]The principles of *Timbral Differentiation* (unique timbral character of a stream) and *Source Location* (spatial separation of sound sources for each stream) are central for stream segregation but are not considered in the paper as we focus on music of a homogeneous timbral character irrespective of source spatialization concerns. We will not refer to the *Toness Principle* and the *Minimum Masking Principle* as these have a less direct influence on voice separation.

Before looking into the way tones are organized 'vertically' and 'horizontally' into coherent 'wholes,' it is important to discuss briefly the principle of *Limited Density*. According to Huron (2001), "if a composer intends to write music in which independent parts are easily distinguished, then the number of concurrent voices or parts ought to be kept to three or fewer" (p. 46). Two issues will be raised in relation to this principle: Firstly, the 'distinguishability' or distinctness of concurrent voices does not imply that one can or should attend to all concurrent voices simultaneously. That is, a voice may be equally distinguishable from other concurrent voices but not necessarily attended to by a listener. Bregman (1990) states in regard to multiple concurrent streams that "we surely cannot pay attention to all these streams at the same time. But existence of a perceptual grouping does not imply that it is being attended to. It is merely available to attention on a continuing basis" (p. 465). Secondly, the number of nominal voices/parts of musical works is often reduced perceptually to a smaller number of auditory streams or perceptual 'voices' via fusion of dependent parts. This is true, for instance, in homophonic or partially homophonic music where sequences of multi-tone sonorities are perceived as individual streams. This way, the density of concurrent streams is reduced, making 'thick' music more accessible to perception.

Vertical Integration

Bregman (1990) explores in depth processes relating to the perceptual integration/segregation of simultaneous auditory components, i.e., how "to partition the set of concurrent components into distinct subsets, and to place them into different streams where they could be used to calculate the spectral properties of distinct sound sources of sound (such as timbre or pitch)" (p. 213). For simplicity, in this study we do not examine the internal structure of musical notes, i.e., fluctuations of harmonics, overtone structure, and so on; we consider notes as internally static events that are characterized by onset, pitch, and duration (as represented in piano-roll notation). In this paper we will focus only on three main aspects of such processes that relate to three principles presented by Huron (2001), namely the principles of *Onset Synchrony*, *Tonal Fusion*, and *Pitch Co-modulation*. The discussion will revolve mainly around the Onset Synchrony Principle since this principle is considered paramount for voice separation and is incorporated in the proposed computational model.

Sounds that are coordinated and evolve synchronously in time tend to be perceived as components of a single auditory event. "Concurrent tones are much more apt to be interpreted by the auditory system as constituents of a single complex sound event when the tones are temporally aligned" (Huron, 2001, p. 39). Concurrent tones that start, evolve, and finish together tend to be grouped together into a single sonority. For instance, in regard to ensemble playing, Bregman (1990) states that "for maximum distinctness, the onset and offset of the notes of the soloist should not be synchronous with those of the rest of the ensemble" (p. 491).

In practical terms, we could state that notes that start concurrently and have the same duration tend to be merged vertically into a single sonority. "Because judgements about sounds tend to be made in the first few hundred milliseconds, the most important aspect of temporal coordination is the synchronization of sound *onsets*" (Huron, 2001, p. 39). Based on the importance of note onsets that determine inter-onset intervals (IOIs), but taking into account also durations, which are less well defined since offsets are perceptually less prominent and more difficult to determine precisely, we can state the following principle:

Synchronous Note Principle: Notes with synchronous onsets and same IOIs (durations) tend to be merged into a single sonority.

This principle relates to Huron's *Onset Synchrony Principle*, i.e., asynchronous note onsets lead to a high degree of perceptual independence of parts (p. 40). However, there is an important distinction between this and the proposed principle: Huron's principle refers only to onsets, not IOIs (or durations). The proposed principle is more specific and gives a more precise account of vertical integration. For instance, in the first example of Figure 7a, the integration of notes with synchronous onsets is much stronger than the integration of the same notes with synchronous onsets in Figure 7b. In the second case, the notes with synchronous onsets have different IOI values (and durations), leading to weaker vertical integration and stronger horizontal sequencing.

Let us examine two musical examples that can be categorized clearly under the labels 'homophony' and 'polyphony.' The two excerpts are drawn from a chorale and a fugue by J. S. Bach (Figures 8 and 9). The chorale is a typical homophonic piece, which is primarily perceived as a single stream that consists of a melody and accompanying harmony (internal voices are very difficult to follow independently—the bass line is not in the primary focus of attention and is an integral part of the harmonic progression). On the contrary, the fugue is a typical polyphonic piece, which is primarily perceived as four independent concurrent streams. In these examples,

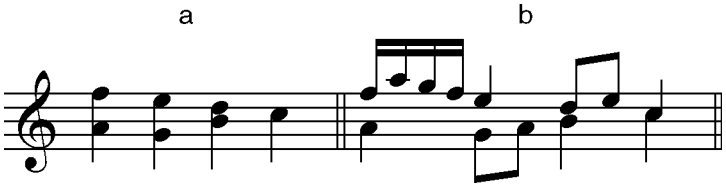

FIGURE 7. Integration of notes with synchronous onsets and same IOIs (a) is much stronger than the integration of the same notes with just synchronous onsets (b).

notes that are vertically integrated are illustrated by quadrangles in which the two parallel vertical sides indicate synchronous note onsets (i.e., synchronous IOIs). It is clear that in the case of polyphony such shapes are sparse, whereas they are abundant in the case of a homophonic texture. Especially in the case of polyphony (Figure 9), notes with synchronous onsets *and* IOIs are much fewer (5 pairs of notes out of a total of 49 notes) than just notes with synchronous onsets (36 pairs out of 49 notes).

A second important factor for vertical integration of tones relates to the *Principle of Tonal Fusion*: The perceptual independence of concurrent tones is weakened when they are separated by intervals (in decreasing order: unisons, octaves, perfect fifths . . .) that promote tonal fusion (Huron, 2001, p. 19). The fusion between synchronous notes is strongest when notes are in unison, very strong when separated by an octave, strong when separated by a perfect fifth and progressively weaker when separated by other intervals. This principle suggests that concurrent pitches are integrated depending on the degree of tonal fusion implied by interval type rather than mere pitch proximity.

In the example of Figure 10, measures 12-14 present a 3-part homophonic passage. All of the computational models presented in the next section would extract 3 voices, except Kilian and Hoos's (2002) model that may split the passage into 2 streams corresponding to the two staves of the score (the left hand notes are placed in the same stream due to pitch proximity). The model proposed by Temperley (2001) traces 3 voices, but he

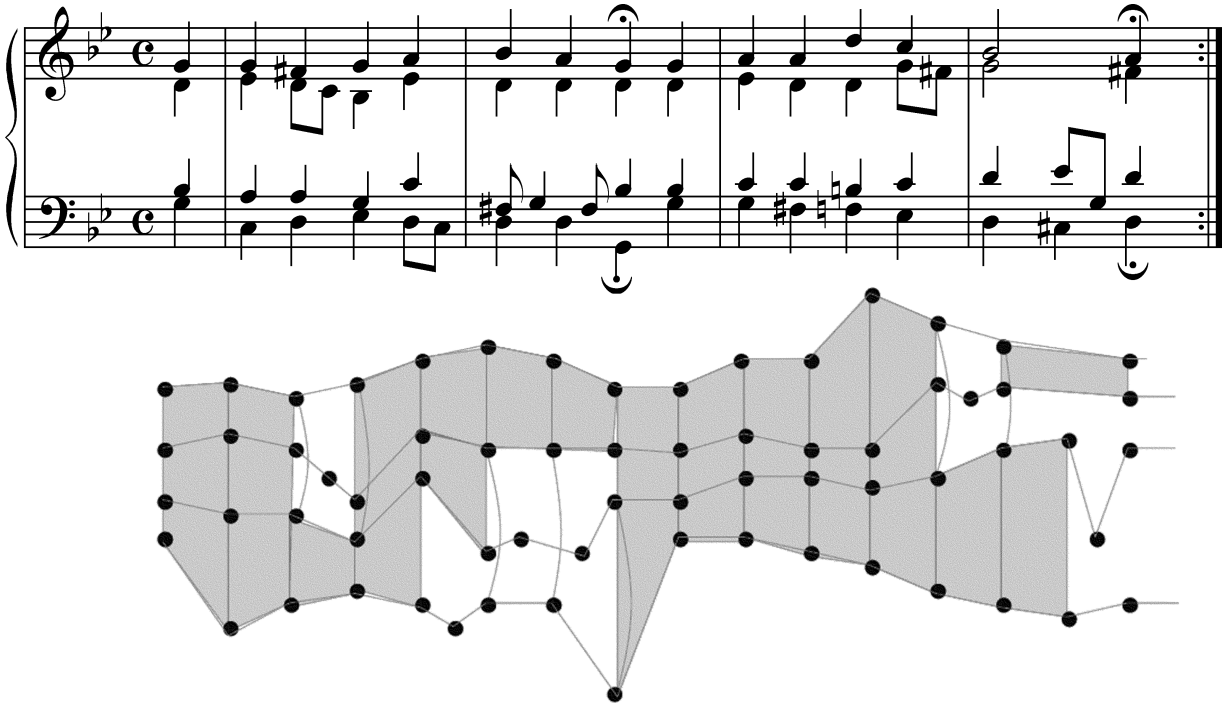

FIGURE 8. First four measures from J. S. Bach's Chorale 73 ('Herr Jesus Christ, du höchstes Gut') as a traditional score and as 'piano-roll' (without durations) with quadrangles illustrating synchronous notes (the two parallel vertical sides of each quadrangle indicate synchronous note onsets).

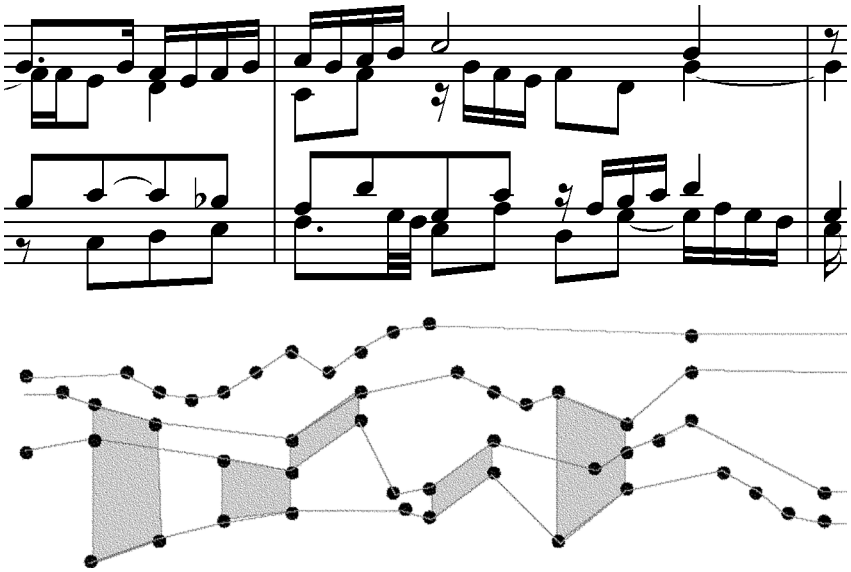

FIGURE 9. Excerpt (mm. 4-5) from J. S. Bach's Fugue 1 in C major, Well-Tempered Clavier, Book 1 as a traditional score and as 'piano-roll' (without durations) with quadrangles illustrating synchronous notes.

notes that "one might suggest that the two left-hand streams from m. 12 onwards form a single larger stream. Such higher level streams would, of course, contain multiple simultaneous notes" (p. 366). It is herein suggested that a listener perceives this passage at maximum as 2 streams (a single stream is also possible) but not the specific two streams suggested above. The passage can be heard as one upper stream consisting of the right-hand part and the upper left-hand part that move in parallel octaves, and a second stream consisting of the lower left hand part. Tonal fusion, in this instance, is more significant for tone integration than pitch proximity. This principle appears to be (at least partially) in conflict with the pitch proximity principle that has been adopted for vertical integration in the computational model by Kilian and Hoos (see next section).

According to the *Pitch Co-modulation Principle*, "The perceptual union of concurrent tones is encouraged

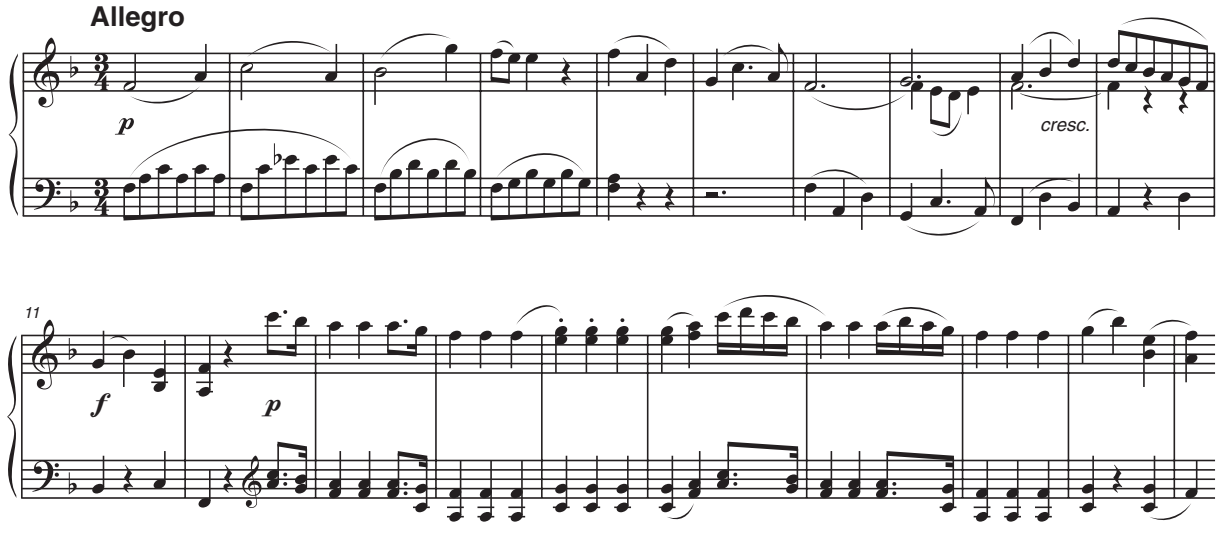

FIGURE 10. Mozart, Sonata K332, I, mm. 1-20.

when pitch motions are positively correlated" (Huron, p. 31). The strongest manifestation of this principle is when notes move in parallel intervals (especially in octaves). This principle implicitly assumes that the onsets of the notes determining the intervals are synchronized. The Pitch Co-modulation Principle can be seen as a special case of the Synchronous Note Principle (or Huron's Onset Synchrony Principle) in the sense that the integration of synchronized note progressions is reinforced when pitch progressions are positively correlated (e.g., moving in parallel octaves, fifths etc.). This principle essentially enables splitting homophonic textures into more than one stream (see, for instance, Figures 5 & 6).

Horizontal Integration

The horizontal integration of musical elements (such as notes or chords) relies primarily on two fundamental principles*:* Temporal and Pitch Proximity. This means that notes close together in terms of time and pitch tend to be integrated perceptually in an auditory stream. These principles are described succinctly by Huron (2001) as follows:

Principle of Temporal Continuity: "In order to evoke strong auditory streams, use continuous or recurring rather than brief or intermittent sound sources. Intermittent sounds should be separated by no more than roughly 800 ms of silence in order to ensure the perception of continuity." (p. 12)

Pitch Proximity Principle: "The coherence of an auditory stream is maintained by close pitch proximity in successive tones within the stream." (p. 24)

These two principles form the basis for all the computational models presented in the next section.

It seems that the temporal continuity principle is a prerequisite for the pitch proximity principle, as the latter requires 'successive tones' in advance of determining proximal pitches. This, however, is only partially true, as it is possible to interpret a tone as belonging to a stream due to pitch proximity, even though it is less temporally continuous than other tones in the context of that stream. Implied polyphony is a clear case where pitch proximity overrides temporal continuity.

Competition Between Vertical and Horizontal Integration

The horizontal integration of tones affects the way tones in vertical sonorities are integrated (and the reverse). Bregman (1990) talks of 'capturing' a tonal component out of a 'mixture.' One of the strongest factors that weakens the vertical links between tones is the appearance of a tone that is proximal to one of the tones of the mixture in terms of both pitch and time. In a sense, there is a competition between the vertical and horizontal principles of auditory grouping. It is exactly this competition that makes it difficult to describe processes of auditory streaming systematically.

In this paper, it is suggested that vertical integration is, in some respect, prior to horizontal sequencing of tones. The idea of capturing a component out of a mixture suggests that the formation of a mixture is anterior to the process of capturing one of its tones into a horizontal stream. This view is different from other (computational) models of 'voice' separation that start off with horizontal organization of streams and then suggest that one could proceed (no model actually does) with vertical integration of streams into higher-level streams that may contain multiple simultaneous tones.

Vertical integration, however, requires estimation of IOIs (according to the Synchronous Note Principle stated above) which means that elementary horizontal streaming is necessary in order to determine which tone onsets define each IOI (durations can be taken into account in this process). The aim is to determine potential note successions rather than full stream separation (the latter is a more complex optimization process). In a sense, vertical integration and horizontal streaming processes evolve concurrently within a certain local context.

The proposed voice separation algorithm (see below) starts by identifying synchronous notes that tend to be merged into single sonorities, and then uses the horizontal streaming principles to break them down into separate streams. This is an optimization process wherein the various perceptual factors compete with each other in order to produce a 'simple' (inasmuch as possible) interpretation of the music in terms of a minimal number of streams (ambiguity, however, should be accommodated).

A preliminary computational prototype that performs voice separation according to some of the principles and processes described above is presented in the next section.

Computational Modeling of 'Voice' Separation

Developing and testing voice separation algorithms allows a systematic exploration of the problem of voice separation. Hypotheses can be set in terms of

formal rules and processes that are thought to be important in the specific musical task, and then these hypotheses can be explored and evaluated in terms of the performance of algorithms that run on specific test data. In practical terms, voice separation algorithms are very useful in computational implementations as they allow preprocessing of musical data, thus opening the way for more efficient and higher quality analytic results; for instance, voice separation algorithms are a critical component of music analysis, music information retrieval, and automatic transcription systems.

In this section, a number of existing voice separation models will first be discussed, and then a novel prototype computational model will be presented.

Related Work

Recently, there have been a number of attempts to model computationally the segregation of polyphonic music into separate 'voices' (e.g., Cambouropoulos, 2000; Chew & Wu, 2004; Kilian & Hoos, 2002; Kirlin & Utgoff 2005; Madsen & Widmer, 2006; Marsden, 1992; Szeto & Wong, 2003; Temperley, 2001).[8] These models differ in many ways but share two fundamental assumptions:

1. 'Voice' is taken to mean a *mono*phonic sequence of successive nonoverlapping musical tones (an exception is the model by Kilian and Hoos, which is discussed further below).
2. The underlying perceptual principles that organize tones in voices are the principles of temporal and pitch proximity (cf. Huron's Temporal Continuity and Pitch Proximity principles).

In essence, these models attempt to determine a minimal number of lines/voices, such that each line consists of successions of tones that are maximally proximal in the temporal and pitch dimensions. A distance metric (primarily in regards to pitch and time proximity) is established between each pair of tones within a certain time window, and then an optimization process attempts to find a solution that minimizes the distances within each voice, keeping the number of voices to a minimum (usually equal to the maximum number of notes in the largest chord). These models assume that a voice is a succession of individual nonoverlapping tones (sharing of tones between voices or crossing of voices is forbidden or discouraged).

For instance, Temperley (2001) proposes a number of preference rules that suggest that large leaps (Pitch Proximity Rule) and rests (White Square Rule) should be avoided in streams, the number of streams should be minimized (New Stream Rule), common tones shared between voices should be avoided (Collision Rule), and the top voice should be minimally fragmented (Top Voice Rule); the maximum number of voices and weight of each rule is user-defined. Cambouropoulos (2000) assumes that tones within streams should be maximally proximal in terms of pitch and time, that the number of voices should be kept to a minimum, and that voices should not cross; the maximum number of streams is equal to the number of notes in the largest chord. Chew and Wu (2004) base their algorithm on the assumption that tones in the same voice should be contiguous and proximal in pitch, and that voice-crossing should be avoided; the maximum number of voices is equal to the number of notes in the largest chord. Szeto and Wong (2003) model stream segregation as a clustering problem based on the assumption that a stream is essentially a cluster since it is a group of events sharing similar pitch and time attributes (i.e., proximal in the temporal and pitch dimensions); the algorithm determines automatically the number of streams/clusters. All of these voice separation algorithms assume that a voice is a monophonic successions of tones.

The perceptual view of voice adopted in this study, which allows multi-tone simultaneities in a single 'voice,' is the most significant difference of the proposed model to the other existing models. In the examples of Figure 11, all existing algorithms (see exception regarding Kilian and Hoos's algorithm below) that are based on purely monophonic definitions of voice would find two voices in the second example (Figure 11b) and three voices in the third example (Figure 11c). It is clear that such voices are not independent voices and do not have a life of their own; it makes more musical sense to consider the notes in each example as a single coherent whole (a unified harmonic sequence/accompaniment). The algorithm proposed in this paper determines that in all three examples we have a single voice/stream.

The voice separation model by Kilian and Hoos (2002) differs from the above models in that it allows entire chords to be assigned to a single 'voice,' i.e., two or more synchronous notes may be considered as

[8]Stacey Davis (2001, 2006) provides a very simple analytical system that determines where transitions occur between implied voices within a monophonic instrumental line.

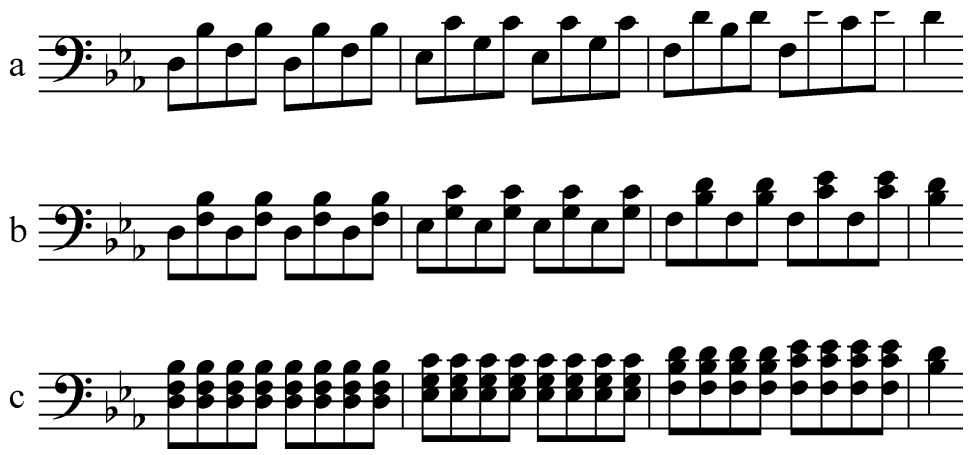

FIGURE 11. Number of voices: in terms of literal monophonic voices all existing computational models will determine in the three examples one, two, and three voices, respectively. In terms of harmonic voices, all examples can be understood as comprising three voices (triadic harmony). In terms of perceptual voices/streams, each example is perceived as a single auditory stream (proposed algorithm).

belonging to one stream.[9] The model is based on a dynamic programming approach. It partitions a piece into slices; contiguous slices contain at least two nonoverlapping notes. A cost function is calculated by summing penalty values for features that promote segregation, such as large pitch intervals, rests/gaps, and note overlap between successive notes, and large pitch intervals and onset asynchrony within chords. Within each slice the notes are separated into streams by minimizing this cost function. The user can adjust the penalty values in order to give different prominence values to the various segregation features, thus leading to a different separation of voices. The maximum number of voices is user-defined or defined automatically by the number of notes in the largest chord.

The aim of the algorithm is to find "a range of voice separations that can be seen as reasonable solutions in the context of different types of score notation" (Kilian & Hoos, 2002, p. 39). The pragmatic goal of the algorithm is the derivation of reasonable score notation—not perceptually meaningful voices. The algorithm is based on perceptual principles, but the results are not necessarily perceptually valid (e.g., a 4-part homophonic piece may be 'forced' to split into two musical staves that do not correspond to perceptually pertinent streams). The algorithm does not discover automatically the number of independent musical 'voices' in a given excerpt; if the user has not manually defined the maximum number of voices, the algorithm automatically sets the maximum number equal to the maximum number of co-sounding notes—in this case the algorithm becomes similar to all other algorithms presented above.

Kilian and Hoos's (2002) model allows multiple synchronous or overlapping tones in a single stream based on pitch and temporal proximity. However, there are two problems with the way this idea is integrated in the model. Firstly, simple pitch and temporal proximity are not sufficient for perceptually pertinent 'vertical' integration. For instance, Kilian and Hoos's model can separate a 4-part fugue into two 'streams' based on temporal and pitch proximity, but these two 'streams' are not perceptual streams but, rather, a convenient way to divide notes into two staves. In perceptual terms, tones merge when they have 'same' onsets and durations (see next section); overlapping tones with different onsets and durations do not merge (there exist,

[9]At this stage, we should additionally mention Gjerdingen's (1994) and McCabe & Denham's (1997) models, which relate to stream segregation but are not considered to be directly 'voice separation' algorithms as their output is not an explicit organization of notes into voices/streams (their model cannot directly be tested against annotated musical data sets). Gjerdingen's model is based on an analogy with apparent motion in vision. In the model each tone of a musical piece has an activation field that influences neighboring tones at a similar pitch. The activation fields of all the tones sum up forming a two-dimensional hill-like activation map; tracing the local maxima in the time dimension on this map produces pitch traces that may be interpreted as streams. Synchronous notes that are proximal in terms of pitch may be merged into a single stream. In this sense, Gjerdingen's model allows concurrent events to be integrated into a single stream based on pitch proximity. This model partially captures the perceptual phenomenon of the greater importance of outer voices.

however, special cases where this happens—not discussed in this paper). Secondly, synchronous notes that are separated by a small pitch interval are not in general more likely to be fused than tones further apart. For instance, tones an octave apart are strongly fused whereas tones a 2nd apart are less likely to be fused. The perceptual factor of tonal fusion is not taken into account by the model.

Kilian and Hoos's (2002) model is pioneering in the sense that multi-note sonorities within single voices are allowed; their model, however, is apt to give results that are erroneous in terms of auditory stream segregation, as this is not the goal of the algorithm.

A New Computational Approach

In this section, a new voice separation algorithm is presented. This algorithm has been developed as a means to explore more systematically the ideas set forth in the earlier sections of this paper; it is not a mature ready-to-use application, but rather an exploratory prototype that requires further development. The proposed prototype is not directly comparable to other algorithms as its underlying definition of 'voice' is different and has a different aim. In this respect, it cannot be compared on a same dataset with other models and one cannot claim that it is better or worse than other algorithms.

The current version of this algorithm is based on three of the principles outlined above: the *Synchronous Note Principle*, the *Temporal Continuity Principle*, and the *Pitch Proximity Principle.* Incorporating the Tonal Fusion and the Pitch Co-Modulation Principles is part of ongoing research.

A very informal description of the algorithm is given below with a view to highlighting its main underlying principles and functions. The details of the actual implementation are presented in (Karydis et al., 2007).

The Proposed Voice Integration/Segregation Algorithm

The proposed prototype accepts as input a musical surface in symbolic form (quantized MIDI) and outputs the number of detected musical voices/streams. At present, the algorithm is applied to quantized musical data (symbolic scores converted to MIDI); expressively performed musical data (e.g., expressive MIDI) require quantization before being fed into the algorithm. The appropriate number of streams is determined automatically by the algorithm and can be lower than the maximum number of notes of the largest chord.

The *VISA* algorithm moves in a stepwise fashion through the input sequence of musical events (individual notes or concurrent note sonorities). Let the entire musical piece be represented as a list L of notes that are sorted according to their onset times. A sweep line, starting from the beginning of L, proceeds to the next onset time in L. The set of notes that have onsets equal to the position of the sweep line is denoted as sweep line set (SLS).

For a set of concurrent notes at a given point, we have to determine when to merge them according to the *Synchronous Note Principle*. Because it is possible that synchronous notes may belong to different voices, we need a way to decide if such merging should be applied. For each SLS, the algorithm examines a certain musical context (window) around them. If inside the window, most co-sounding notes have different onsets or offsets, then it is most likely that we have polyphonic texture (independent monophonic voices), so occasional synchronous notes should not be merged—each note is considered to be a singleton cluster. If most notes are concurrent (same onsets and IOIs) implying a homophonic texture, then they should be merged—concurrent notes form a cluster. This way, each SLS is split into a number of note clusters.[10]

For each SLS in the piece, we have a set of previously detected voices (V) and the current set of note clusters (C). Between every detected voice of V and each note cluster of C, we draw an edge to which we assign a cost. The cost function calculates the cost of assigning each cluster to each voice according to the *Temporal Continuity Principle* and the *Pitch Proximity Principle* (clusters and voices that are closer together in time and pitch receive a lower cost). Notes that overlap receive a cost value equal to infinity.

A dynamic programming technique finds the best matching (lowest cost) in the bipartite graph between previous voices and current clusters. If voices are fewer than clusters, then one or more voices may be terminated. If clusters are fewer than voices, then new voices may appear. The matching process, additionally, takes into account two constraints. The first one is that voice crossing should be avoided. Therefore a suboptimal solution in terms of cost may be required that avoids voice crossing. The second one is that the top voice

[10]At the present stage, the window size *w* and homophony/polyphony threshold *T* have been determined manually (same for all the data) by finding values that give optimal results for the selected test data set.

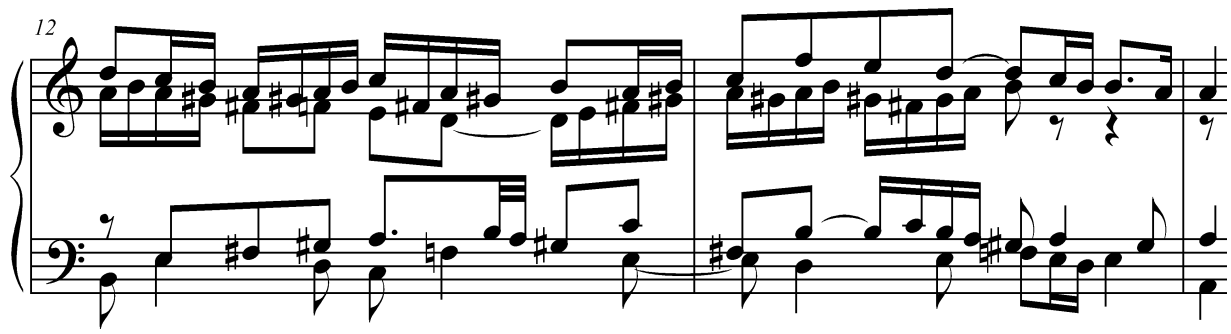

FIGURE 12. The proposed algorithm performs voice separation correctly in this excerpt from the Fugue No. 1 in C major, WTCI, BWV 846 by J. S. Bach, except for the last five notes of the upper voice which are assigned to the second voice (see text).

should be minimally fragmented (Top Voice Rule by Temperley, 2001). This is handled by adding a penalty to the cost of a matching that does not fulfil this rule; to find the matching with the minimal cost, a cluster may be split into subclusters, so that one can be assigned to the top voice.

Results

The proposed algorithm has been tested on a small set of musical works for piano. Eight pieces with clearly defined streams/voices act as ground truth for testing the performance of the algorithm. The first four pieces are two fugues from the first book of the Well-Tempered Clavier by J. S. Bach (Fugue No. 1 in C major, BWV 846, and Fugue No. 14 in F# major, BWV 859) and two inventions by J. S. Bach (Invention No. 1 in C Major, BWV 772, Invention No. 13 in A minor, BWV 784); these polyphonic works consist of independent monophonic voices. Two mazurkas (Op. 7, No. 5 and Op. 67, No. 4) and a waltz (Op. 69, No. 2) by F. Chopin consist of a melody (upper staff) and accompanying harmony (lower staff). Finally, the Harmony Club Waltz by S. Joplin has two parallel homophonic streams (chordal 'voices') that correspond to the two piano staves. See musical excerpts in Figures 12, 13, and 14.

In this pilot study, the aim was to examine whether a single algorithm can be applied to two very different types of music (i.e., pure polyphonic music and music containing clear homophonic textures). All the parameters of the algorithm are the same for all eight pieces; the number of streams/voices is determined automatically (not set manually). It should be noted that for the four pieces by Chopin and Joplin all other voice separation algorithms would automatically determine at least four different voices (up to eight voices) that cannot be considered as independent perceptual sequences.

Annotated datasets for musical streams/voices (as 'voices' are defined in this paper) do not exist. A small dataset was therefore selected for which it is assumed that musicologists/musical analysts would unreservedly agree on the number of independent musical

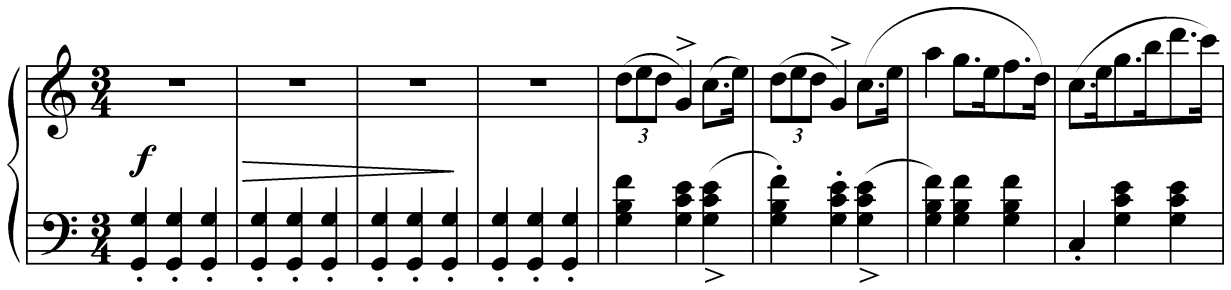

FIGURE 13. In the opening of the Mazurka, Op. 7, No. 5 by F. Chopin, the proposed algorithm correctly detects one voice (low octaves) and then switches automatically to two voices (melody and accompaniment).

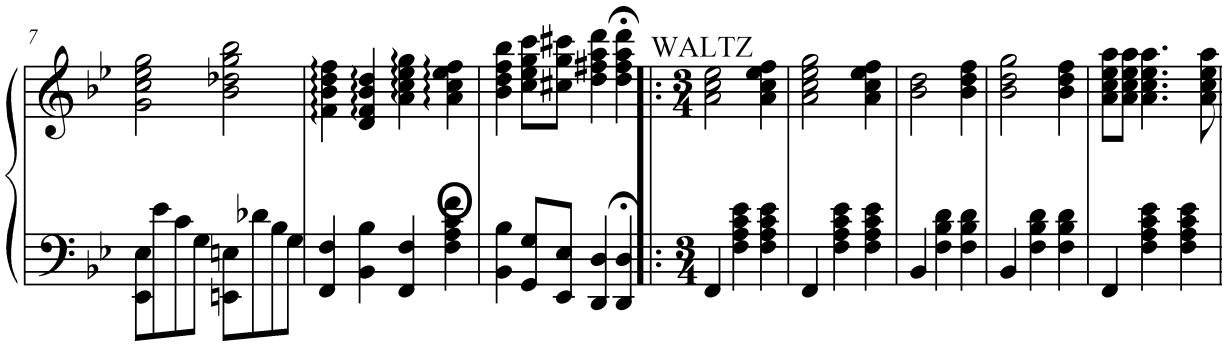

FIGURE 14. Two independent chordal streams/voices are correctly determined by the algorithm in this excerpt from the Harmony Club Waltz by S. Joplin—mistake is indicated by the circled note (see text).

streams in each piece.[11] The small size of the test dataset is problematic as overfitting can occur. However, the fact that the pieces in the dataset have been selected from contrasting musical styles (polyphonic vs homophonic) guarantees that the dataset is not biased towards one homogeneous music dataset. Working on a small dataset has enabled both quantitative and qualitative evaluation of the results (all the results have been analyzed note-by-note and mistakes have been categorized in types of problems—see below). A larger dataset, however, is assembled in order to run larger scale tests.

The evaluation metric used is the precision of the obtained result (percentage of notes clustered correctly). For the previously described musical dataset, Table 1 shows the results. The effectiveness of the proposed methodology is evident by the high precision rates achieved for the eight pieces.

The aforementioned results were examined in detail in order to understand the kinds of mistakes produced by the algorithm. Most of these problems are, in a sense, expected and cannot be solved when taking into account merely pitch and temporal distances between notes.

The majority of wrong results were given in cases where the number of voices change, and erroneous connections are introduced primarily due to pitch proximity (for instance, in Figure 12 the algorithm erroneously 'gives' the last five notes of the upper voice to the second voice simply because the first of these notes is closer by a semitone to the last note of the second voice). Kilian and Hoos (2002) address this same problem claiming that, in essence, it is unsolvable at the note level ("It seems that when only considering the notes . . . there is no reason why another separation should be preferred," p. 45).

A second kind of problem involves voice crossing. Since voice crossing is disallowed by the algorithm, notes at points (in the Bach fugues) where voices cross are assigned to wrong voices.

A third type of mistake relates to the breaking of vertically merged notes into subsonorities and allocating these to different voices. In this case, the breaking point in the sonority may be misplaced (see, for instance, circled note in Figure 14).

The success rate of the algorithm on this small dataset, comprising both homophonic and polyphonic works, is remarkable, and it shows the potential of the whole approach. The proposed computational prototype, however, requires further testing on a large annotated dataset and further development in order to become more robust.

TABLE 1. Results in Terms of Precision for the Dataset.

Musical Work	*Precision*
J. S. Bach, Fugue No.1 in C major, BWV846	92.3%
J. S. Bach, Fugue No.14 in F# major, BWV859	95.5%
J. S. Bach, Invention No.1 in C Major, BWV 772	99.3%
J. S. Bach, Invention No.13 in A Min, BWV 784	96.4%
F. Chopin, Mazurka, Op.7, No. 5	100%
F. Chopin, Mazurka in A Minor, Op. 67, No. 4	85.0%
F. Chopin, Waltz in B Minor, Op. 69, No. 2	90.3%
S. Joplin, Harmony Club Waltz	98.1%

[11]These independent 'voices' correspond to separate spines in the kern format; all test pieces have been obtained from KernScores http://kern.humdrum.org

Conclusions

In this paper the notions of voice and auditory stream have been examined, and an attempt has been made to clarify the various meanings, especially of the term 'voice,' within various musicological, psychological, and computational contexts. It is suggested that if voice is understood as a musicological parallel to the concept of auditory stream, then multi-note sonorities should be allowed within individual voices. For the sake of argument, the current paper has taken a rather strong view on the correspondence between musical stream and voice, bringing these notions to a nearly one-to-one relationship. The aim was to raise a productive and fruitful debate on the issue rather than to provide definitive answers.

Various perceptual principles pertaining to auditory stream integration/segregation have been briefly examined (primarily in relation to Huron's exposition of these principles), as they form the basis of attempts to formalize voice separation processes. It has been suggested that the two principles of temporal and pitch proximity are insufficient to form the basis of the general problem of voice/stream separation, and that they have to be complemented primarily by the Synchronous Note Principle (and also by the Tonal Fusion and Pitch Co-modulation Principles).

It is proposed that, if a primarily perception-oriented view of voice is adopted, voice separation models should take into account vertical integration of multitone sonorities along with horizontal stream segregation. A general musical voice/stream separation model that can function not only in cases where music comprises a fixed number of monophonic voices, but in the general case where homophonic/polyphonic/heterophonic elements are mixed together, should consider incorporation of co-sounding events in a single 'voice.' Whether the term 'voice separation' is most appropriate to describe such general models is open to further discussion.

The proposed voice separation algorithm incorporates the two principles of temporal and pitch proximity, and additionally, the Synchronous Note Principle. Allowing both horizontal and vertical integration enables the algorithm to perform well not only in polyphonic music that has a fixed number of 'monophonic' lines, but in the general case where both polyphonic and homophonic elements are mixed together. We have shown in the above preliminary experiment that a single algorithm, with the same parameters, can achieve good performance in diverse musical textures (homophonic and polyphonic) in terms of identifying perceptually relevant voices/streams. Ongoing research involves testing the algorithm on a much larger database, and incorporating additional principles such as the Tonal Fusion and Pitch Co-modulation Principles.

Author Note

Correspondence concerning this article should be addressed to Emilios Cambouropoulos, Department of Music Studies, Aristotle University of Thessaloniki, University Campus of Thermi, 54124, Thessaloniki, Greece; E-MAIL: emilios@mus.auth.gr

References

BREGMAN, A. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge, MA: MIT Press.

CAMBOUROPOULOS, E. (2000, July/August). *From MIDI to traditional musical notation*. Paper presented at the American Association for Artificial Intelligence Workshop on Artificial Intelligence and Music: Towards Formal Models of Composition, Performance and Analysis, Austin, Texas.

CHEW, E., & WU, X. (2004). Separating voices in polyphonic music: A contig mapping approach. In U. Kock Wiil (Ed.), *Computer Music Modeling and Retrieval: Second International Symposium* (pp. 1-20). Berlin: Springer.

DANN, E. (1968). *Heinrich Biber and the seventeenth century violin*. Unpublished doctoral dissertation, Columbia University.

DAVIS, S. (2001). *Implied polyphony in the unaccompanied string works of J. S. Bach: Analysis, perception and performance*. Unpublished doctoral dissertation, Columbia University.

DAVIS, S. (2006). Implied polyphony in the solo string works of J. S. Bach: A case of perceptual relevance of structural expression. *Music Perception*, *23*, 423-446.

DRABKIN, W. (2007). Part-writing [voice-leading]. *The New Grove Dictionary of Music and Musicians*. Retrieved September 2, 2007, from: http://www.grovemusic.com/

GJERDINGEN, R.O. (1994). Apparent motion in music? *Music Perception*, *9*, 135-154.

HURON, D. (1989). Voice denumerability in polyphonic music of homogeneous timbres. *Music Perception, 6*, 361-38.

HURON, D. (2001). Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception*, *19*, 1-64.

KARYDIS, I., NANOPOULOS, A., PAPADOPOULOS, A.N., & CAMBOUROPOULOS, E., (2007) VISA: The voice integration/segregation algorithm. In S. Dixon, D. Bainbridge, & R. Typke

(Eds.), *Proceedings of the Eighth International Conference on Music Information Retrieval* (pp. 445-448). Vienna: Austrian Computer Society.

KILIAN, J., & HOOS, H. (2002). Voice separation: A local optimisation approach. In M. Fingerhut (Ed.), *Proceedings of the Third International Conference on Music Information Retrieval (ISMIR'02)* (pp. 39-46). Paris: IRCAM—Centre Pompidou

KIRLIN, P. B., & UTGOFF, P. E. (2005). VoiSe: Learning to segregate voices in explicit and implicit polyphony. In J. D. Reiss & G. A. Wiggins (Eds.), *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR'05)* pp. 552-557. London: Queen Mary, University of London.

MADSEN, S. T., & WIDMER, G. (2006). Separating voices in MIDI. In R. Dannenberg, K. Lemström, & A. Tindale (Eds.), *Proceedings of the Seventh International Conference on Music Information Retrieval* (pp. 57-60). Victoria, Canada: University of Victoria.

MARSDEN, A. (1992). Modeling the perception of musical voices: A case study in rule-based systems. In A. Marsden & A. Pople (Eds.), *Computer representations and models in music* (pp. 239-263). London: Academic Press.

MCADAMS, S., & BREGMAN, A. S. (1979). Hearing musical streams. *Computer Music Journal, 3*, 26-43.

MCCABE, S. L., & DENHAM, M. J. (1997). A model of auditory streaming. *Journal of the Acoustical Society of America, 101,* 1611-1621.

PISTON, W. (1991). *Harmony.* London: Victor Gollancz Ltd.

SCHOENBERG, A. (1963). *Preliminary exercises in counterpoint.* London: Faber and Faber Ltd.

SWAIN, J. P. (2002). *Harmonic rhythm: Analysis and interpretation.* New York: Oxford University Press.

SZETO, W. M., & WONG, M. H. (2003). A stream segregation algorithm for polyphonic music databases. In B. C. Desai & W. Ng (Eds.), *Proceedings of the Seventh International Database Engineering and Applications Symposium* (pp. 130-138). Hong Kong, China: IEEE Computer Society.

TEMPERLEY, D. (2001). *The cognition of basic musical structures.* Cambridge, MA: MIT Press.