‘Voice’ separation: theoretical, perceptual and computational perspectives

Emilios Cambouropoulos
Department of Music Studies
Aristotle University of Thessaloniki
emilios@mus.auth.gr

ABSTRACT

The notions of ‘voice’, as well as, homophony and polyphony, are thought to be well understood by musicians. 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. The paper presents different views of what ‘voice’ means and how the problem of voice separation can be described systematically, with a view to understanding the problem better and developing a systematic perceptually-based description of the cognitive task of segregating ‘voices’ in music. Vague (or even contradicting) treatments of this issue will be presented. Elements of a systematic theory that can be implemented as a computer program are also proposed.

WHAT IS A VOICE?

It appears that the term ‘voice’ has different meanings for different research fields (traditional musicology, music cognition, computational musicology). Recently, there have been a number of attempts (e.g. Temperley, 2001; Cambouropoulos 2000; Kilian & Hoos 2002; Szeto and Wong, 2003; Chew & Wu 2004; Kirlin & Utgoff 2005) to model computationally the segregation of polyphonic music into separate ‘voices’. Much of this research is influenced by empirical studies in music perception (e.g. Bregman, 1990; Huron 2001), as well as by more traditional musicological concepts such as melody, counterpoint, voice-leading and so on.

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 (Fig. 1b). Finally, this succession of tones can even be separated into three different voices (Fig. 1c) if the implied triadic harmony is taken into account (Fig. 1d).

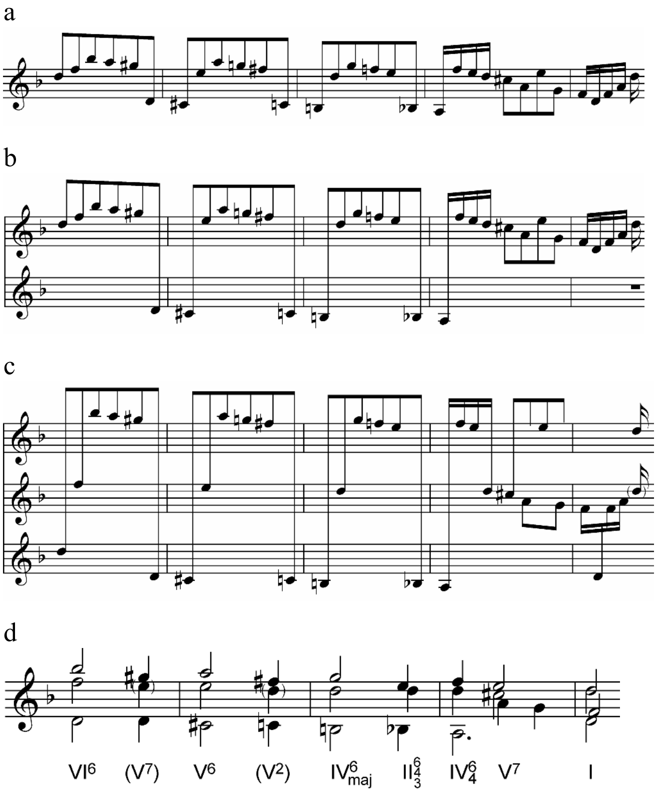

Figure 1 Measures 33-36 from Bach’s Chaconne for solo violin from the D Minor Partita (BWV 1004) presented as: a) one voice (solo violin), b) two voices (perceived implied polyphony), or c) three voices following the implied triadic harmonic structure (harmonic reduction[1] presented in 1d).

For this musical passage, we can see three different ways in which ‘voice’ may be understood: a) literally

Proceedings of the 9th International Conference on Music Perception & Cognition (ICMPC9).

[1] Harmonic analysis provided by music theorist Costas Tsougras.

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 where these notions are incongruent. Before discussing 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 form the Greek word ‘phōnē’ (φωνή) which means ‘voice’ (of humans or animals) or even ‘the sound of musical instruments’.[2] In this sense, the term voice is used primarily to refer to the sound sequences produced by different monodic 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 (see section 5). 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).
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 organisation 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 (such as the one presented in Figure 1d). If such a harmony is perceivable, one could hypothesise that a listener is capable of organising 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).

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

In all of the above descriptions, implicit is the assumption that ‘voice’ is a *mono*phonic sequence of successive non-overlapping 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. In this paper, the plain term ‘voice’ will refer to this meaning.

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 ‘voice’). 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 listen to the individual constituent lines at all (especially the ‘inner’ middle voice, or even the lower voice). The perceptual principles

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

[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, ch.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).

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.

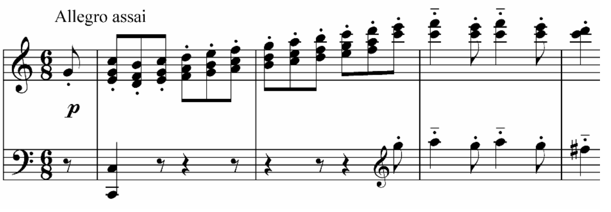

Figure 2 Opening measures of Beethoven's Sonata Op.2, No.3. Should this passage be understood as determining three parallel voices or a single chordal voice?

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 implicit polyphony 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 organises 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.

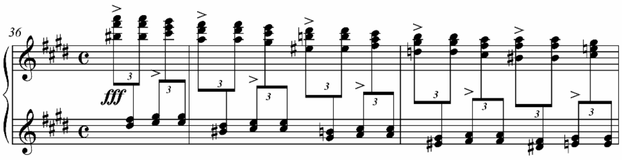

Figure 3 Excerpt from Rachmaninov's Prelude Op.3, No.2 (mm.36-38).

In the next section, the relationship between the notion of voice and stream will be discussed.

VOICE AND STREAM

David Huron's acclaimed 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.

Most of the current research in voice separation modelling refers to this award-wining paper as giving the perceptual framework which may form the basis for the development of actual computational systems. Researchers take, as a starting point, one or more of the proposed perceptual principles and then develop computational models that are tested on a set of multi-voice musical works. The discussion below is aimed not at refuting Huron's claims in regard to voice-leading and perceptual principles but rather to show that the relation between voice and stream can be more complicated depending on what is meant by each of these terms.

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 a 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 non-musical 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 perceives individual inner parts in timbrally undifferentiated 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 in general that do not always determine independent sequences of tones.

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 proceeds proposing two additional perceptual goals in the case of homopnony (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 (1980/2006) 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, p2). Independency of melodic lines is supported both by rhythmic and melodic factors. If one insists in giving primacy on one of these two factors, it is more plausible that rhythmic independency 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 non-traditional 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 independency. Consider, for instance, the two musical examples in Figures 8 and 4. In the first example (Figure 8), 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 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 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 4), 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 marginally traditional, or, actually, 'non-traditional' as parallel or similar movement between voices for relatively long periods is usually avoided (see, also, example of two streams of block chords in Figure 5; this is a case of homophony where common-practice considerations of voice-leading are disregarded).

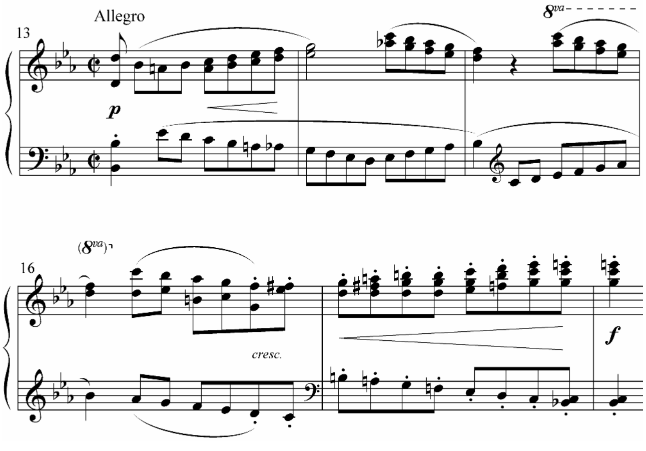

Figure 4 Excerpt from Beethoven's Sonata Op.13 *Les Adieu* (mm. 13-18)

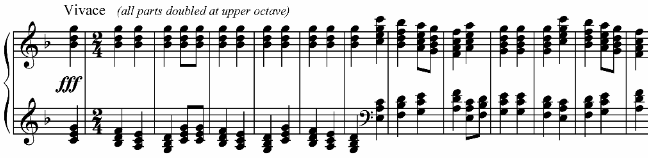

Figure 5 Two streams of 'block chords' in Stravinsky's, Petrushka, Tableau I. Piston (1991, p.488) regards this excerpt as two lines thickened by block chords.

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 independent streams may be perceived (as in the examples presented in figures 4 and 5). 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 fibre itself (except the melody).

The core hypothesis underlying David Huron's study can be summarised 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 to, 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:

1. 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 'twisting' perceptual principles (by means of choosing only the ones that are considered appropriate) to 'fit' the pre-accepted 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 lead to perceptually independent voices).

2. The approach taken by Huron gives 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 approach is unwarranted. Some traditional voice-leading rules may correctly be thought of 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 music taste and convention (e.g. parallel fifths[5] are not necessarily 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 music 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 scope to developing a formal system (implementable on a computer) that is capable to achieve perceptually meaningful 'voice separation' (not necessarily to discover 'voices' indicated in the score by the composer). In order not to confuse the commonly understood musicological notion of voice with the extended notion of perceptual 'voice' proposed in this paper, we will use the less confusing term 'stream' which will be considered, for practical reasons, as equivalent to 'perceptually independent voice consisting of single or multi-note sonorities'.

The aim of the above discussion is to show that the notions of voice and stream are not always congruent. A voice (as a monophonic sequence of tones) is not always perceived as a musical stream, and conversely, a musical stream is not always a voice. Voice-leading is not related in a one-to-one manner to musical stream segregation. 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 stream segregation (for instance, by means of computational modelling).

COMPUTATIONAL MODELS OF VOICE SEPARATION

'Voice' separation algorithms are very useful in computational implementations as they allow preprocessing of musical data opening thus 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 melodic voices and/or 'voices' consisting of multi-note sonorities can assist more sophisticated processing within the voices (rather than across voices). For instance, if one wants to identify

[5] 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).

musical works that contain a certain melodic pattern, this pattern should be found not spread across different parts (perceptually implausible) neither 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.

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

1. 'Voice' is taken to mean a *mono*phonic sequence of successive non-overlapping musical tones (exception is the model by Kilian and Hoos which will be discussed further below)
2. The underlying perceptual principles that organise 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 optimisation process attempts to find a solution that minimises 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 non-overlapping 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 large leaps (Pitch Proximity Rule) and rests (White Square Rule) should be avoided in streams, the number of streams should be minimised (New Stream Rule) and common tones shared between voices should be avoided (Collision Rule)[6] – 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 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. more than one synchronous notes may be considered as belonging to one stream.[7] The model partitions a piece into slices; each slice contains at least a certain number of notes different from adjacent slices. 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 within chords. Within each slice the notes are separated into streams by minimising this cost function. The user can adjust the penalty values in order to give different prominence values to the various segregation features leading thus 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.

This algorithm introduces the pitch proximity principle as an integrating factor not only in the 'horizontal' dimension between successive notes, but also 'vertically' between synchronous notes. If the user has defined a relatively low maximal number of voices (e.g. 2 voices) and the number of notes in a given slice is greater than this number (e.g. 4 notes), then these notes are separated into sub-chords based on the pitch proximity factor (e.g. into one 3-note chord and a single note, or two 2-note chords) – for instance, a Bach chorale presented in Kilian and Hoos's paper is split into four individual voices if the system determines automatically the number voices (equal to the number of notes in the largest chord) or into two streams corresponding roughly to piano staff notation if the user sets the maximum number of voices to two.

Kilian and Hoos's model allows multiple synchronous tones in a single stream. However, there are two serious problems with the way this idea is integrated in the model. Firstly, pitch and temporal proximity are not sufficient for '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

[6] An additional fifth rule takes care that the top voice is minimally fragmented (Top Voice Rule).

[7] 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 herein to be directly 'voice separation' algorithms as their output is not an explicit organisation 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 which influences neighbouring 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 in a single stream based on pitch proximity; this model partially captures the perceptual phenomenon of the greater importance of outer voices.

perceptual terms, tones merge when they have 'same' onsets, not simply when they are proximal (cf. Huron's Onset Synchrony Principle). Two voices may be moving closely together in terms of pitch and still be perceived as independent because of different rhythmic patterns. 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 (cf. Huron's Tonal Fusion Principle). These two important perceptual factors are not taken into account by the model; the model is therefore doomed to make serious mistakes in terms of stream integration/segregation.

PERCEPTUAL PRINCIPLES FOR 'VOICE' SEPARATION

In this section, fundamental principles of perceptual organisation 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 may 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 10 principles that cover all major aspects of stream integration/segregation, we will use a number of these principles[8] as the starting-point of our exploration (it is assumed, however, that the reader is acquainted with these principles):

The principles of *Onset Synchrony* and *Limited Density* are not considered as optional auxiliary principles that shape music in perceptually distinctive ways (giving rise to different musical genres), but as fundamental perceptual principles that enable a listener to 'break down' the flow of musical tones into independent streams. These streams 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 organises musical tones into coherent auditory streams or 'voices').

Before looking into the way tones are organised '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 to 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). In this paper we will focus only on two aspects of such processes that relate to two principles presented by Huron (2001), namely the principles of *Onset Synchrony* and *Tonal Fusion*. For simplicity, in this study we do not examine the internal structure of musical notes, i.e. flunctuations 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).

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)

[8] *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: Asynchronous note onsets lead to a high degree of perceptual independence of parts. (p.40)

Principle of Limited Density: Concurrent parts ought to be kept to three or fewer, if they are to be easily distinguished. (p.46)

In practical terms, we could state that notes that start concurrently and have 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 of the importance of note onsets that determine 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.*

In Figure 6, the integration of notes with synchronous onsets in the first example (Fig. 6a) is much stronger than in the second example (Fig. 6b). In the second case, the notes with synchronous onsets have different IOI values (and durations); this leads to weaker vertical integration and to stronger horizontal sequencing.

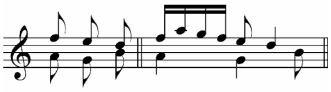

Figure 6a & b

Since defining IOIs (Inter-Onset Intervals) requires elementary note streaming information (i.e. which note onset follows the current note), we will re-examine this issue in the next section (horizontal integration of notes).

This principle relates to Huron's *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) However, there is an important distinction between this and the proposed principle: Huron's principle is applied *after* streams have been established whereas the proposed principle is applied concurrently with horizontal streaming processes (see below for details). In this sense, the proposed principle is more fundamental as it is an integral part of the streaming process itself rather than an auxiliary post-streaming effect.

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; 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 (2002).

In the example of Figure 7, measures 12-14 present a 3-part homophonic passage. Most of the above computational models would extract 3 voices, except Kilian and Hoos's 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). Temperley's model traces 3 voices, but he 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.' (Temperley, 2001, 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.

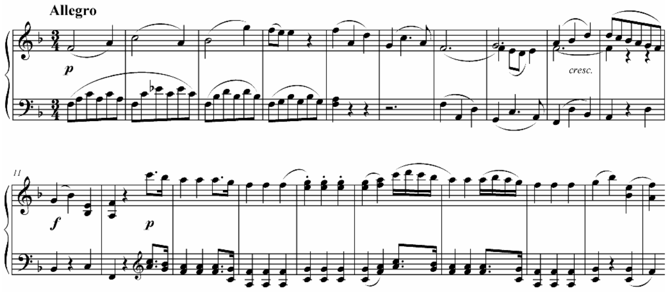

Figure 7 Mozart, Sonata K332,I, mm.1-20

According to the *Pitch Co-modulation Principle:* 'The perceptual union of concurrent tones is encouraged 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 synchronised. 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 synchronised 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 in more than one stream (see, for instance, examples 5 & 6). It is surprising that Huron considers the more specialised principle as a primary principle and the more general/fundamental one an auxiliary principle.

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 800ms of silence in order to ensure the perception of continuity.' (Huron, 2001, 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)

It seems that the temporal continuity principle is a prerequisite for the pitch proximity principle as the latter requires 'successive tones' in advance to 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 clear case where pitch proximity overrides temporal continuity.

Vertical vs. 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 systematically processes of auditory streaming.

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 in contrast to most models of 'voice' separation that start off with horizontal organization of streams and then proceed (or at least suggest that one should proceed) with vertical integration of streams into higher-level streams that may contain multiple simultaneous tones.

However, vertical integration 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 which is a more complex optimisation process.

Let us examine two musical examples that can clearly be categorised under the labels 'homophony' and 'polyphony' - the two excerpts are drawn from a chorale and a fugue by J.S.Bach (Figs 8 & 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 hardly perceptible – 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 notes that are vertically integrated are illustrated by quadrangles in which the two parallel vertical sides indicate synchronous note onsets. It is clear that in the case of polyphony such shapes are sparse, whereas they are abundant in the case of a homophonic texture.

It is suggested, that a voice separation algorithm should start by identifying synchronous notes that tend to be merged into single sonorities and then use 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' (as much as this is possible) interpretation of the music in terms of a minimal number of streams (ambiguity, however, should be accommodated).

It is beyond the scope of this paper to present a full working computational model that performs 'voice' separation according to the above discussion (this is currently investigated and should lead to a working model in the near future) The aim of the paper is to highlight the various aspects of both the meaning itself of voice and stream, and the perceptual factors that affect the exact way voice separation processes work.

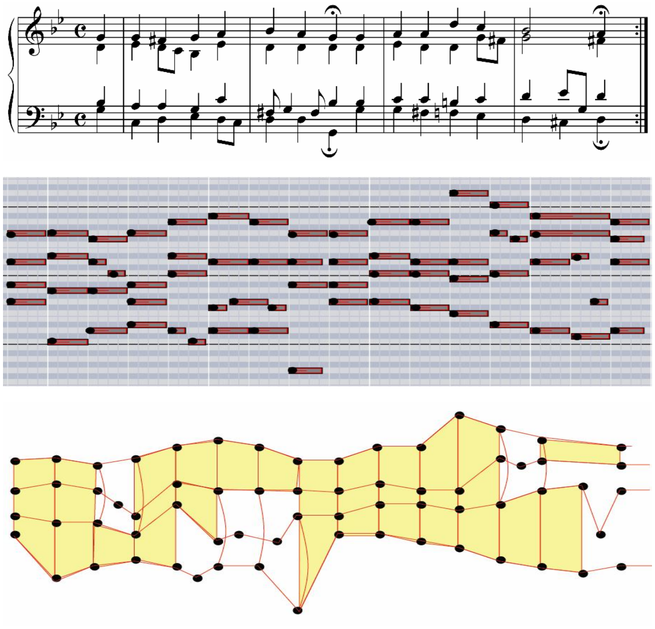

Figure 8 First four measures from J.S.Bach's Chorale 73 ('Herr Jesus Christ, du höchstes Gut') as a traditional score, as piano-roll notation and as piano-roll (without durations) with quadrangles illustrating synchronous notes (the two parallel vertical sides of each quadrangle indicate synchronous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

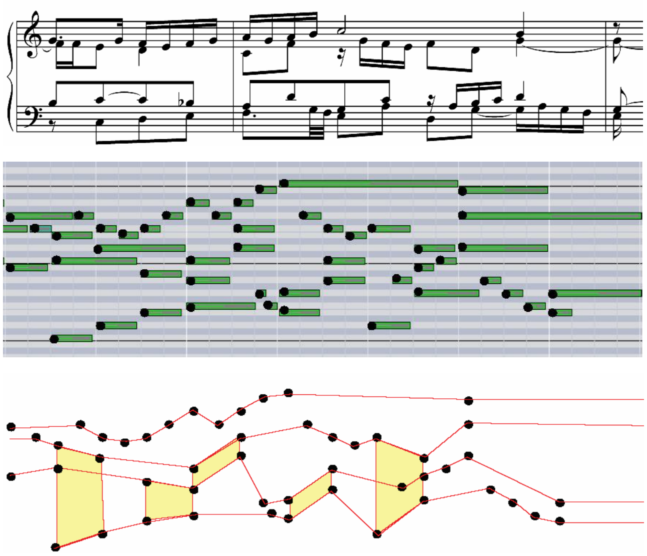

Figure 9 Excerpt (mm.4-5) from J.S.Bach's Fugue 1 in C major, Well-Tempered Clavier, Book 1 as a traditional score, as piano-roll notation and as piano-roll (without durations) with quadrangles illustrating synchronous notes (the two parallel vertical sides of each quadrangle indicate synchronous note onsets).

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 allowed within individual 'voices'.

The 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 are often the basis of attempts to formalise 'voice' separation processes. It has been suggested that the two principles of temporal and pitch proximity are insufficient to form the basis of 'voice' 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 a first step in voice separation is identifying synchronous note sonorities and then breaking these into sub-sonorities incorporated in horizontal streams or 'voices'. This proposal is in direct contrast with most computational systems that start by finding first horizontal 'voices' and then merging these into higher level 'voices' (actually, the latter step has not been implemented by any of the aforementioned computational models).

References

Bregman, A (1990) *Auditory Scene Analysis: The Perceptual Organisation of Sound.* MIT Press, Cambridge (Ma).

Cambouropoulos, E. (2000) From MIDI to Traditional Musical Notation. In *Proceedings of the AAAI Workshop on Artificial Intelligence and Musid: Towards Formal Models of Composition, Performance and Analysis*, July 3 - Aug. 3, Austin Texas.

Chew, E. and Wu, X. (2004) Seperating voices in polyphonic music: A contig mapping approach. In *Computer Music Modeling and Retrieval: Second International Symposium* (CMMR 2004), pp. 1-20.

Dann, E. (1968) *Heinrich Biber and the Seventeenth Century Violin*. Ph.D. Thesis, Columbia University.

Huron, D. (2001) Tone and Voice: A Derivation of the Rules of Voice-Leading from Perceptual Principles. *Music Perception*, 19(1):1-64.

Gjerdingen, R.O. (1994) Apparent Motion in Music? *Music Perception*, 9(2):135-154.

Kilian j. and Hoos H. (2002) Voice Separation: A Local Optimisation Approach. In *Proceedings of the Third International Conference on Music Information Retrieval* (ISMIR 2002), pp.39-46.

Kirlin, P.B. and Utgoff, P.E. (2005) VoiSe: Learning to Segregate Voices in Explicit and Implicit Polyphony. In *Proceedings of the Sixth International Conference on Music Information Retrieval* (ISMIR 2005), Queen Mary, University of London (pp. 552-557).

Marsden, A. (1992) Modelling the Perception of Musical Voices: a Case Study in Rule-based Systems. In *Computer Representations and Models in Music*, Marsden, A. and Pople, A. (eds), Academic Press, London.

McCabe, S.L. and Denham, M.J. (1997) A Model of Auditory Streaming. *The Journal of the Acoustical Society of America*, 101(3):1611-1621.

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

Temperley, D. (2001) *The Cognition of Basic Musical Structures*. The MIT Press, Cambridge (Ma).

Schoenberg, A. (1963) Preliminary Exercises in Counterpoint. Faber and Faber Ltd, London.

Szeto, W.M. and Wong, M.H. (2003) A Steam Segregation Algorithm for Polyphonic Music Databases. In *Proceedings of the Seventh International Database Engineering and Applications Symposium* (IDEAS'03).