AUDITORY SCENE ANALYSIS AND SOUND SOURCE COHERENCE AS A FRAME FOR THE PERCEPTUAL STUDY OF ELECTROACOUSTIC MUSIC LANGUAGE

Blas Payri, José Luis Miralles Bono
Universidad Politécnica de Valencia, Campus de Gandía,
46730 Grao de Gandía, Spain,
bpayri@har.upv.es, josmibo@posgrado.upv.es

Abstract

This paper proposes a method for the perceptual study of sound language and sound material in electroacoustic music based on ASA (auditory scene analysis), using contextual listening and sound source coherence as the main perceptual task. Claiming that out-of-context listening of sounds to describe their features - either timbre or morphological features – is not sufficient to understand their musical value, the paper proposes to combine it with contextual listening. The research paradigm is ASA, using the notions of stream segregation and abstract source coherence perception. The acousmographe is a sufficient tool for ASA based analysis of musical sequences, and the experiments can combine a top-down approach (understanding the features perceived in preexisting musical works) and a bottom-up approach – creating musical sequences where features vary in a controlled way. This allows a real musical listening that can be compared with non contextual listening to understand which features really are musically salient.

INTRODUCTION

The most frequent experimental frame for the study of the perception of timbre has its roots in the pioneering work of [Grey, 1977] and [Wessel, 1979]. In these experiments, timbre is defined as what distinguishes sounds with same pitch, same loudness and same duration. The sound material is made of (synthetic) instrumental samples, calibrated in pitch and loudness. Listeners are asked to rate the global similarity between samples, and this leads to a "timbre distance". For example, [McAdams, 1989] develops the notions of timbre intervals and timbre distances that could be used as form-bearing elements. This experimental frame has permitted to accumulate a set of perceptual dimensions and their acoustic explanations, but meets several drawbacks. One drawback is that the "timbre space" that arises from this experiments is very dependent on the set of sounds [Donnadieu *et al.*, 1994] [McAdams, 1993], and that the variety of sounds used do not reflect the diversity of instrumental or electroacoustic music sounds. The normalization of sounds in pitch, duration and loudness reduces the complexity of real sounds. And, maybe the most important drawback, the fact that sounds are listened out of any context biases the perception of humans if we want to understand the use of timbre or other sound features in music language or other sound based languages like soundscapes or audiovisual works' soundtracks. Indeed, any notion of language implies the relation of elements within a context. It is important to analyze these

drawbacks and understand how we can perceive sound features and their usability in the "language" component of music language.

EXPERIMENTS WITH VOCAL TIMBRE

Why vocal timbre

As we want to study in language implications of timbre perception, we have chosen to study speech timbre as speech is what best defines language. Vocal timbre and musical timbre research have some common ground in the experimental design and history: when dealing with the issue of voice timbre, most perceptual studies use as sound material voice samples from different speakers. This samples may be sustained vowels ([Walden *et al*., 1978], [Murry *et al*., 1977], [Kempster *et al*., 1991]), a read word or sentence [Murry *et al*., 1978], [Murry *et al*., 1980], [Fagel and van Herpt, 1983], [Kreiman *et al*., 1992], or more rarely, spontaneous speech. Most often these utterances are produced "at a comfortable speaking level" in non-communicative settings. The target of these studies is to discover and explain the axes of the perceptual space for the voice. We should point out that in most cases the sound material is only concerned with inter-speaker variability (each sample comes from a different speaker) and not at all with intra-speaker variability (for example, different pitches, loudness, intention or prosody produced by the same speaker). We have the same limitations as discussed above: researchers tend to associate one sample with one speaker the same way that musical timbre associates one instrument to one sample, sound production is normalized and the sounds are out of context.

Taking these restrictions in account, we needed to define our experimental settings for vocal timbre perception in a way that would allow the comparison of contextual and non-contextual listening.

Sound material

We have chosen as basic material recordings from the italian version of the EUROM european project. 20 utterances were chosen, representing 10 male and 10 female speakers. The sentence uttered was the Italian word "seicento cinquantotto" (meaning "six-hundred and fifty-eight"), read by 20 Italian speakers. All the recordings have been performed in a silent environment, with the same position from the speaker to the microphone, using the same recording (high quality) material and sampling frequency (20kHz). The speakers were all native Italians and had no voice pathologies. Their ages ranged from 20 to 50 years, and according to the authors of the soundbase, they had been chosen to represent vocal timbre diversity.

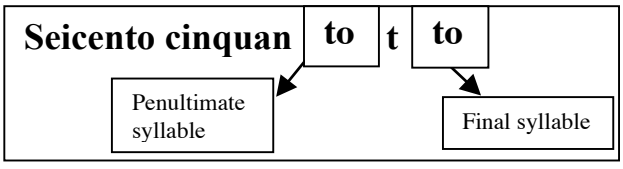

Figure 1: Whole utterance and extracted syllables used as a sound material

Using this material we have built four different sound material types for the purposes of our research:

1. **Original context**: Entire utterance "seicento cinquantotto", for which we had 20 samples. This material allows a minimum contextual listening.
2. **Out-of-context**: A set of 30 syllables coming from the previous utterances. The syllables were the last two "to" syllables of the "cinquantotto" as shown in **Erreur ! Source du renvoi introuvable.**. Thus, for most of the 20 speakers we had two samples: the final and the penultimate syllable of the utterance.
3. **Mixed context**: we replaced the penultimate syllable "to" in an entire utterance described in point 1 with the equivalent syllable of another speaker as shown in figure 2. We insist on the fact that the different segments are simply concatenated, no crossfade or pitch change was performed. As we performed all possible combinations of the 20 entire utterances with the 20 penultimate "to" syllables, we obtained a total of 400 combinations (20 original context plus 380 multi-speaker mixtures).
4. **Parameter modification with mixed content:** we took one of the 20 original utterances, (female speaker) in which we inserted the original syllable "to" plus 19 unmodified syllables "to". We also modified the F0 of the inserted syllables in order to match the F0 of the replaced syllable (19 mixes). Finally we modified the F0 of the original syllable "to" with 10 degrees of variation (changes by semitones). We had a total of 48 combinations.

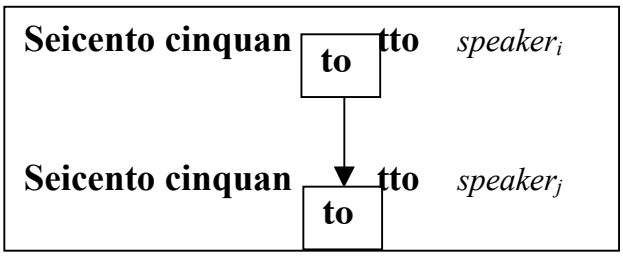

Figure 2: Replacement the penultimate tonic "to" syllable in a speaker's utterance by another equivalent syllable from another speaker's utterance.

Global similarity task experiments

Listeners had to classify the samples according to global similarity, which is a task of holistic listening similar to the experiments in musical timbre perception and speaker timbre perception of the literature. Subjects used an interface made for this purpose where they were presented with the whole set of samples, subjects could move each icon corresponding to each sound sample to classify the sounds. Once they had finished their classification they validated the results and described each of the resulting classes they had created.

These experiments were made with the "original context" and the "out of context" sound material described above, that is, with the unmodified 20 utterances "seicento cinquantotto" and then, in a second period, with the isolated "to" syllables.

Axis rating experiments

Using the same sound material as the global similarity task, listeners had to rate each sample (utterance, syllable) along several axes with opposite adjectives: male-female, feminine-non feminine, masculine-non masculine, pleasant-unpleasant, high pitch-low pitch, tense-relaxed and estimated age.

Source coherence experiments

This experiments were performed with the "mixed context" and "parameter modification with mixed context" described above. The subjects had to listen to each sound and answer the question "Has this sentence been pronounced by one speaker or is there a mixture of different speakers?". The answer was binary, and the subjects could hear as many times as wished the sounds. The listening was done in a sound-isolated cabin with a high quality loud speaker, or with high quality headphones. The sounds were presented by pages of 40 sounds. The proportion of mixed and original utterances was balanced: 20 mixed and non-mixed utterances per page, and as there were only 20 original utterances for 380 mixed utterances, the original utterances were repeated in each page. Each mixed utterance was present in only one page (chosen randomly for each subject), and the order of mixed and non-mixed utterances was randomized for each page. We should note that repeating the original utterances in each page and having a half mixed-half non-mixed proportion led to a higher discrimination and rejection of the mixed utterances.

Results: 1) contextual versus non-contextual listening

From the answers of the subjects on the global similarity task for syllables and utterances we deduced that the syllables and the utterances were not completely classified according to the same criteria. One important result was that the syllables were not classified according to the speaker, but instead according to their position in the sentence: no syllable in final position was classified with syllables in penultimate position. We deduced from this that the changes in voice quality (pitch, loudness, tension…) caused by the prosody for two contiguous syllables are perceptually more relevant than some inter-speaker voice quality difference. It indicated that taken out of context, syllables may sound different from each other but they can blend together smoothly within the utterance of the speaker.

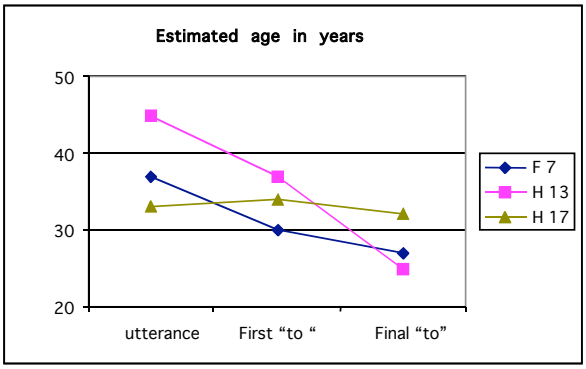

Figure 3: Age ratings for the whole utterance and extracted syllables

The results of axis rating experiments show also a significant difference between the ratings of the entire utterance and the isolated syllables listened out of context. Listeners can rate quite differently the perceived age of the whole utterance, the first and the second extracted "to" syllables as displayed for example in figure 3. More surprisingly, a very salient and robust feature as perceived gender displayed differences between the syllables and the whole utterance from which they were extracted, as shown in figure 4. These results are important for our purpose as they show that timbre qualities may be rated differently for isolated syllables, and then integrate in the context of an utterance resulting in different timbre qualities ratings. This Gestalt analysis where the whole is different from the parts has deep

consequences in timbre perception, indicating that in the context of a musical sequence, the features rated out of context may not apply.

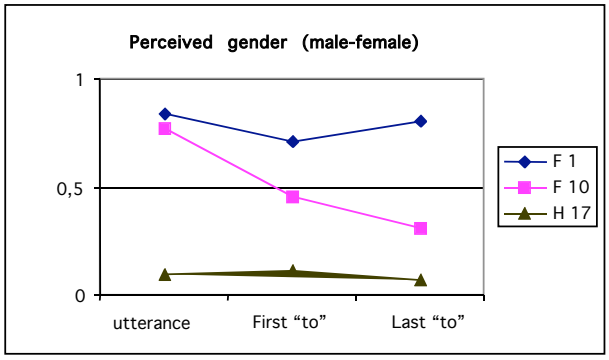

Figure 4: Gender ratings for the whole utterance and extracted syllables

Results: 2) contextual distance and similarity distances

From the answers of the subjects we derive a contextual distance: it measures, for a given utterance, the ratio of the answers "it is a mixed utterance" over the total number of answers. If this distance is high (close to 1) it means that the utterance is perceived as mixed, if the distance is low (close to 0) then the utterance is perceived as coming from a single speaker: this means that the substitution of the syllable is highly acceptable. After eliminating aberrant answers, (some subjects rejected even the original utterances), we used 333 pages of answers, which gives a mean of 18 answers per mixed utterance and 333 answers per original one.

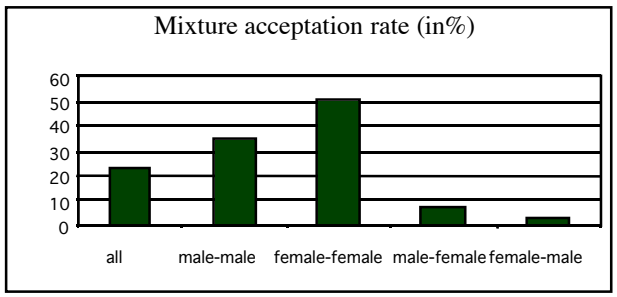

Figure 5: Acceptation rate of the mixed context utterances where one syllable of another speaker was introduced grouped by gender of the utterance and the inserted syllable speakers.

As displayed in figure 5, many mixtures were accepted (more listeners rated it as "same speaker" than "several speakers") specially when restricting the mixtures to same-gender speakers, which is obvious as pitch is the most salient perceptual feature. We performed an INDSCAL analysis on the contextual analysis. The two first dimensions accounted for 67% of the variance. The most obvious result from is the correlation between the first dimension and the pitch-gender axis. This dimension is also correlated with the INDSCAL dimension calculated from the similarity distance for utterances and syllables. It means that for a syllable to substitute another syllable, it is necessary that they have near pitch values and that the speakers they originate from have also near mean pitch values as can be seen in figure 6.

The second INDSCAL dimension correlated highly with the third INDSCAL dimension from the similarity distance for the syllables: this dimension is so far unexplained by an acoustic axis.

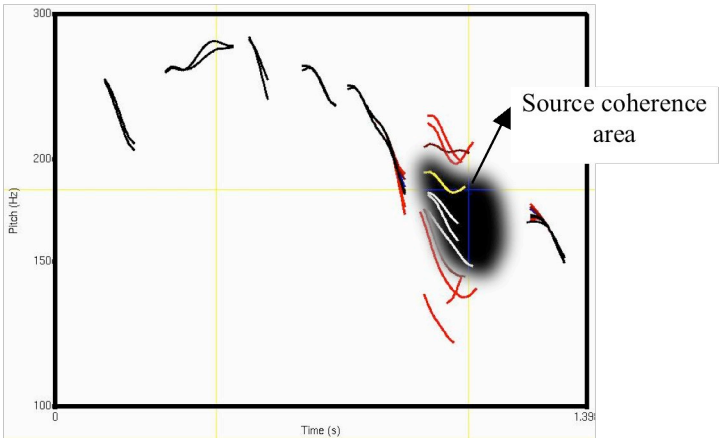

Figure6: Pitch contours of the different mixtures of a given utterance and the 20 "to" syllables. In red, the syllables that were rejected, in black on white and white on black, the syllables that were accepted. The dark area represents the zone were all or most syllables are accepted within the utterance context.

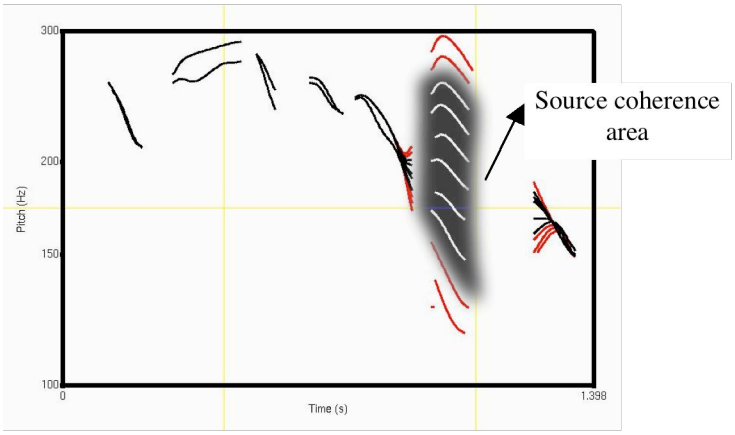

Figure 7: Pitch contours of an utterance with the original "to" syllable modified in F0. In red, the syllables that were rejected, in black on white and white on black, the syllables that were accepted. The dark area represents the zone were all or most syllables are accepted within the utterance context

The results of the contextual distance for the experiment where the sound material included syllables modified in F0 show that the acceptation rate was significantly higher for different values of F0 as can be seen in figure 7. This means that when timbre features are "very similar" (that is when the syllable modified is the original syllable in this utterance context), pitch variations can be greater. So we have a new analysis on the contextual distance.

CONCLUSIONS

The main conclusions of our experiments is that the context has a great influence in the perception of timbre, and perceptual features vary depending that an element is heard isolated or in the context of a sentence. We can derive from that the notion of contextual distance, which measures the fact that a sound heard in the temporal context of our sounds is

perceived as "belonging" or not to the context sequence. In other words if we have stream segregation or not, using the notions of Auditory Scene Analysis [Bregman, 1990]. We can extrapolate the results to instrumental sound perception and electroacoustic music perception. To understand the usability in musical language of a given feature, it is important to listen to it within a context. For instance, we can synthesize a given instrument, changing the tension (or velocity of attack) during a musical phrase. Each note may be perceptually very different, with high and low tension, but in the context of evolving sounds, the global perception of the musical phrase may be coherent or not, and the global tension respond to different rules than the local tension of each note. Stimuli can be any succession of a small number of events, for example, the repetition of a sound object that is progressively transformed. The listeners would then judge whether they hear a single stream, or whether the transformed sound object is heard as a new element belonging to a second stream.

In electroacoustic music, [Camilleri and Smalley, 1998] point out that the perceptual approach has taught us that it is not viable practice to separate an account of sonic materials and musical structure from signification. We propose then a gestalt approach [Leman and Schneider, 1997] to know what elements are perceptual salient in music language. We can analyse the sound material of real works, using the tools developed for electroacoustics. The acousmographe [Delalande, 1998] [Couprie, 2004] allows the subject to listen to a sound sequence: the subject performs the analysis by setting graphic symbols on each sound event, as shown in figure 8. The shape and colour of the symbols will indicate whether the events correspond to the same or different sound material type. This task, with sounds heard in their musical context can yield a contextual distance between sound events.

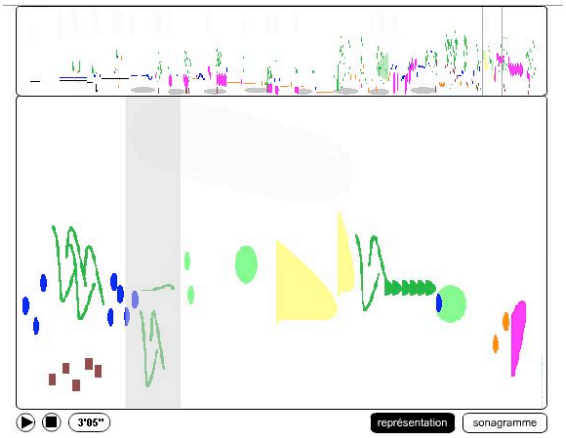

Figure 8: Partial acousmographe analysis on Bayle's "Oiseau moqueur" [Couprie, 1999].

The contextual listening proposed by the acousmographe can be complemented with out-of-context listening tasks, in which the sound material is the different sound events originating from the musical work. The combination of the two experiments will give information on how the musical context changes the similarities between sound events, that is to say, how the composer forces the listener to focus on certain aspects and oppositions of the sound material by "composing" (etymologically, "laying together") the sounds.

REFERENCES

Bregman, A. S. (1990) *Auditory scene analysis*. The MIT press.

Camilleri, L.; Smalley, D. (**1998**) "The analysis of electroacoustic music: introduction", *Journal of New Music Research,* 27, No 1-2

Couprie, P. (1999) "Three Analysis Models for L'oiseau moqueur, one of the Trois rêves d'oiseau by François Bayle", *Organised Sound*, 4 (1), pp.3-14.

Couprie, P. (2004) "Graphical Representation: An analytical publication tool for electroacoustic music", *Organised Sound*, 9 (1), pp.109-113.

Delalande, F. (1998), "Music analysis and reception behaviours: *Sommeil* by Pierre Henry", *Journal of New Music Research,* vol. 27, No 1-2

Donnadieu, S., McAdams, S., Winsberg, S. (1994), "Context effects in timbre space", *3rd Intnl. Conf. on Music Perception and Cognition: ESCOM*, Liège, 311-312.

Fagel, W. P. F.; van Herpt, L. W. W. (**1983**) "Analysis of the perceptual qualities of dutch speakers' voice". *Speech Communication*, 2, pp.315-326.

Grey, J.M. (**1977**), "Multidimensional perceptual scaling of musical timbres", *Journal of the Acoustical Society of America,* 61, pp.1270-1277

Kempster, G.; Kistler, D.; Hillenbrand, J. (**1991**) "Multidimensional scaling analysis of dysphonia in two speaker groups", *Journal of Speech and Hearing Research*, 34, pp.534-543.

Kreiman, J.; Gerratt, B.R.; Precoda, K.; Berke, G.S. (**1992**) "Individual differences in voice quality perception". *Journal of Speech and Hearing Research*, 35, pp.512-520.

Leman, M.; Schneider, A. (**1997**), "Origin and Nature of Cognitive and Systematic Musicology: An Introduction", *Music, gestalt and computing: studies in cognitive and systematic musicology,* Springer, Lecture Notes in Artificial Intelligence 1317

McAdams, S. (**1989**) "Psychological constraints on form-bearing dimensions in music", *Contemporary Music Review*

McAdams, S. (**1993**), "Recognition of Auditory Sound Sources and Events", *Thinking in sound: the cognitive psychology of human audition*, Oxford University Press

Murry, T.; Singh, S.; Sargent, M. (**1977**) "Multidimensional classification of abnormal voice qualities", *Journal of the Acoustic Society of America*, 61(6).

Murry, T.; Singh, S. (**1978**). "Multidimensional classification of normal voice qualities", *Journal of the Acoustic Society of America*, 64, pp.81-87

Murry, T.; Singh, S.; Sargent, M. (**1980**) "Multidimensional analysis of male and female voices", *Journal of the Acoustic Society of America*, 68(5), pp.1294-1300

Walden, B.; Montgomery, A.; Gibeily, G.; Prosek, R.; Schwartz, D. (**1978**) "Correlates of psychological dimensions in talker similarity", *Journal of Speech and Hearing Research*, 21, pp.265-275.

Wessel, D.L. (**1979**) "Timbre space as a musical control structure", *Computer Music Journal,* 3 (2)