Human Speech as a Resource for Music Composition

A Fala Humana como Recurso para a Composição Musical

Bruno Ruviaro
brunoruviaro@gmx.net

Abstract: This research focuses on the sonic aspects of human speech as a source for compositional procedures assisted by computers. We are particularly interested in verifying what kind of musical structures can emerge or be derived from a sample of typical daily-life speech. This paper describes ongoing research into speech analysis and re-synthesis methods designed for musical composition. A theoretical introduction explains briefly the history of electro-acoustic music based on human voice, as well as some basic linguistics concepts related to the nature of speech and language. Conclusions demonstrate some practical results in which human speech functions as the basis of short musical excerpts generated on computer. The use of speech data in instrumental composition is also shown in the conclusions. This research is the basis of the author's present compositional work.

Keywords: Electro-Acoustic, Computer-Aided Composition, Speech

Resumo: O foco desta pesquisa é o estudo de aspectos sonoros da fala humana como fonte de procedimentos composicionais assistidos por computador. Estamos particularmente interessados em verificar que tipo de estruturas musicais podem ser derivadas de amostras de fala cotidiana. Este artigo descreve uma pesquisa em andamento de análise sonora da fala e métodos de re-síntese aplicados à composição musical. Conclusões demonstram alguns resultados práticos nos quais a fala serve de base para a criação de pequenas passagens musicais eletroacústicas. O uso de técnica similar para composição instrumental também é exemplificado.

Palavras-chave: música eletroacústica; composição assistida por computador; fala humana

Introduction

Some expressions have been created in order to categorize certain electro-acoustic compositions which share any of the multiple uses of human voice as their main musical ideas: "text-sound piece", "sprachkomposition", "verbal composition", "hörspiel", among others. The specific approach to the use of human voice in such pieces may vary from a microscopic sonic research into the very essence of a single phoneme to theatrical experiments where the meaning of a text plays a fundamental role.

A number of different uses of known languages and also language simulations can be found in text-sound pieces in general, as well as a broad range of vocal expressions beyond the categories of normal speech or singing. The world of loudspeakers has also contributed to the use of very intimate and otherwise inaudible sounds that one can produce with the vocal apparatus. For example, the use of recorded whispers, subtle moans and breathing (just to cite a few possibilities), isolates those sounds from their private bodily spaces of existencesince usually one can hear them only in very familiar situations at close distances to another body. Such sounds can be put then in different spatial dimensions and presented in the public space of a music concert.

This research focuses on the sound flow of human speech as a source for speculations on musical structures. We are particularly interested in verifying what kind of musical organizations can emerge, or be derived from, a given sample of a typical daily-life conversation.

Looking back in the history of electro-acoustic music, the use of human voice has played an important role in its evolution since the very beginning. The well known pieces Symphonie pour un homme seul (Pierre Schaeffer & Pierre Henry, 1950) Gesang der Jünglinge (Karlheinz Stockhausen, 1955-56), Thema (Ommagio a Joyce) and Visage (Luciano Berio, 1958 and 1961) are good examples. Epitaph für Aikichi Kuboyama (Herbert Eimert ,1960-1962), although not so famous as the others, is another especially interesting composition where a deep research on the borders of music and language was undertaken.

From the last thirty years, we could cite the following examples in order to have a concise overview of the kind of composition we relate with our study: Speech Songs (Charles Dodge ,1973), Requiem (Michel Chion, 1973), Six Fantasies on a Poem by Thomas Campion (Paul Lanksy, 1979) Paul Lansky, The Blind Man (Barry Truax,1979), Mortuos Plango, Vivos Voco (Jonathan Harvey, 1980), PANLaceramento della parola (Ommagio a Trotskij) (Flo Menezes, 1988) and Tongues of Fire (Trevor Wishart, 1994).

Among all those examples we can find close correlations between human voice and electronic sounds; strong links between music and modern poetry and literature; stress of the theatrical side of human voice, either based on a "real" text or on language simulation; importance of a chosen text topic (such as politically engaged themes or religious texts); pioneering speech synthesis research; and creation of new sonic worlds through strict manipulation of voice sounds. These are just a few of the aspects that allow us to see how diverse the approach of human voice can be.

Not in all cases, however, the sound of human voice was used as a source or model for musical organization. The question is: has any aspect of the sound of human voice influenced or determined the structure of the musical composition itself? This provides us with an analytical toolnot a judgment criterionto approach those compositions with which our research is particularly concerned.

In this paper we are going to describe one method of approaching the sound shape of human speech as a main source for generation of musical ideas. At this initial stage of our research, the two main outcomes of this method are computer-generated sounds made through additive synthesis and instrumental compositions based on data extracted from speech.

Speech Analysis/Synthesis Paradigm

Useful Linguistics Concepts

Generally speaking, the linguistic analysis of speech phenomena has some

similarities with our general concept of musical analysis: the main concern is to "divide the continuous sound-flow into a definite number of successive units" (Jakobson, 1956, p. 3)meaningful units, in a higher level, and their minutest constituents, in a lower level. In linguistics, the smallest element endowed with meaning is considered to be the morpheme. Its inner components, which make possible differentiating morphemes from each other, are the phonemes and the distinctive features. Opposition and contrast form the so called "polarity principle", that is to say, the "choice between two terms of an opposition that displays a specific differential property, diverging from the properties of all other oppositions" (Jakobson 1956, p. 4).

The knowledge of the basic categories of small units of spoken sounds is also of great value for the composer interested in a deeper exploration of this field. These categories are vowels, diphthongs, semivowels, nasals, fricatives, plosives, affricates and the whispered consonant H. One of the main musically interesting oppositions here is the one between noise and voiced sounds, which form the basic difference between consonants and vowels[1].

The so-called "supra-segmental" levels of linguistics analysis will be studied in detail during the next steps of this research. By supra-segmental organization we understand the study of language structures above the phonemic level, as for example the mora, the syllable and the foot structures. The study of stress organization is especially relevant to a broader understanding of the prosodic level of human speech. Stress is not directly related to one single physical parameter: changes in pitch and duration have the most influential elements on stress, while loudness has "the least effect on stress perception, despite its intuitive status as the most natural correlate of stress" (Hayes, 1995, p. 6).

Speech Analysis Methods

According to Dodge (1985), two common methods of analysis gained importance since the origin of computer-based analysis of speech: formant tracking and linear predictive coding (LPC).

A formant is a characteristic peak of amplitude in certain frequency regions of the spectrum. It is mainly because of different formant configurations, resulting in different timbres, that we can distinguish one vowel from each other. This is also true for the recognition of timbres in the instrumental domain. "In formant tracking, the analysis transforms the speech signal into a series of short-term spectral descriptions, one for each segment. Each spectrum is then examined in sequence for its principal peaks, or formants, creating a record of the formant frequencies and their levels versus time." (Dodge, 1985, p. 225). The other method, linear predictive coding, is a subtractive analysis/re-synthesis method which

"analyzes [a sound] into a data-reduced form, and re-synthesizes an approximation of it. A prediction algorithm tries to find samples at positions outside a region where one already has samples" (Roads, 1996, p. 200).

The speech analysis currently undertaken for our research is based on a function called Partial Tracking from the software Audiosculpt (IRCAM, Paris). After obtaining a sonogram of the sample under study, the partial tracking function gives us a graphic representation of the partials by means of line segments; these lines are as straight as the partials are steady. In the case of human speech, as we discussed above, the prosodic level is often characterized by continuous changes in pitch within words, thus resulting in curved lines for each partial (Figure 1, left).

We can export this data in text format for subsequent use in our resynthesis. However, the amount of data that would be extracted from all these curved lines for each partial is large, necessitating a correspondingly large computational time in the resynthesis. In order to avoid this problem, we can average all partials so as a single straight line can represent each of them[2]. Audiosculpt gives us this possibility, and then we get the following result:

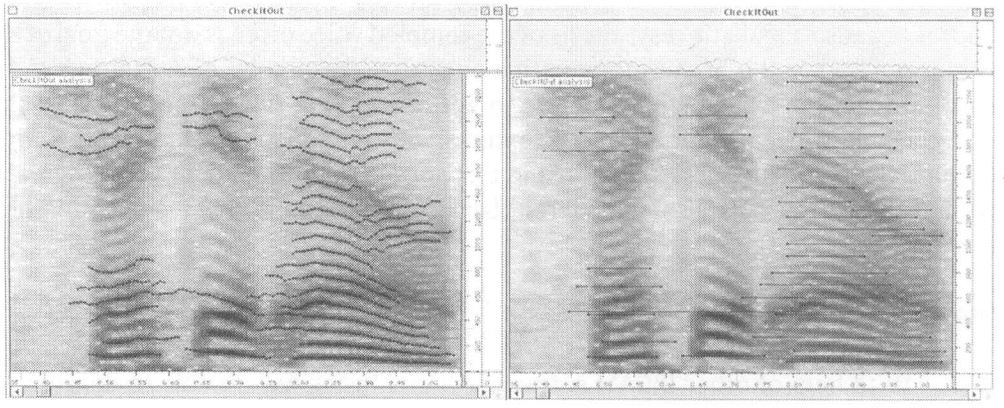

Figure 1. Two sonograms in Audiosculpt. The line segments are the result of a typical partial tracking of a spoken sentence, with lots of small glissandi (on the left). In this example, the words "Check it out" are spoken by a male voice. The same sonogram with averaged partials is shown on the right. Part of the richness of the prosodic features is sacrificed in the averaged analysis.

Once saved into a text file, we have a list of data with the following appearance:

```
(PARTIALS   124
(POINTS     2
0.122       925.909 -29.630
0.383       925.909 -29.630)
(POINTS     2
0.128       2126.556 -33.339
0.232       2126.556 -33.339)
[...]
```

The header indicates the total number of partials. Each sublist contains the starting and ending points of a single straight line (partial), in seconds, and its average frequency (Hz) and amplitude (dB). This list then will be subject to a simple sorting into sublists, one for each individual parameter of the collection of partials: onset, duration, frequency and amplitude.

Speech "Resynthesis"

Perhaps the most interesting step from the musician's viewpoint is that one right before the use of all those data in a "resynthesis" process. We put it between quotes because a realistic resynthesis of the sampled voice often is not the goal of a composer. Rather, it is at this moment that one can make all sorts of alterations and mutations on the available parameters to get results that range from the closest recognizable speech shape to sound textures without apparent relationship with the initial voice. The possibility of using this speech data to feed an instrumental composition project is also considered. Let us now examine these two different approaches.

Csound

With help of other IRCAM software (Open Music), we are able to deal with the raw data in the text files and sort it in a suitable manner to write a Csound score[3]. An enormous variety of complex sound textures can be obtained, resembling the original voice in different levels. In general, as long as the onsets and durations are kept approximately close to the original values, the resulting sound should resemble the temporal contour of the analyzed voice. For example, if the frequencies are radically and irregularly changed but the time values are as in the original, some listeners may still be able to recognize the speech-like rhythms. On the other hand, changing parameters in a combined manner beyond certain limits, sound textures with no resemblance to the human voice can also be obtained. One example is increasing the durations of each partial by 10 or 20 times the original, but keeping frequencies and onsets as in the original. The result will be a kind of long, stretched cloud of sounds in continuous movement. This is because,

as one can imagine, each temporally enlarged partial ends up overlapping with all others several seconds after its onset.

Chord-Seq Module

Instead of managing the speech analysis to write a CSound score, one can also make use of the raw data to develop instrumental compositions. Another Open Music patch is used to convert speech data into a piano score. A few operations must be done on the original list in order to get values in midicents, milisseconds and velocity (amplitude). Open Music module "Chord-Seq" allows us to input these parameters and save a MIDI file. This file can be used as source material for further developments of an instrumental piece based on speech analysis. Some speech-like qualities produce a very special shaping of musical ideas in the instrumental domain.

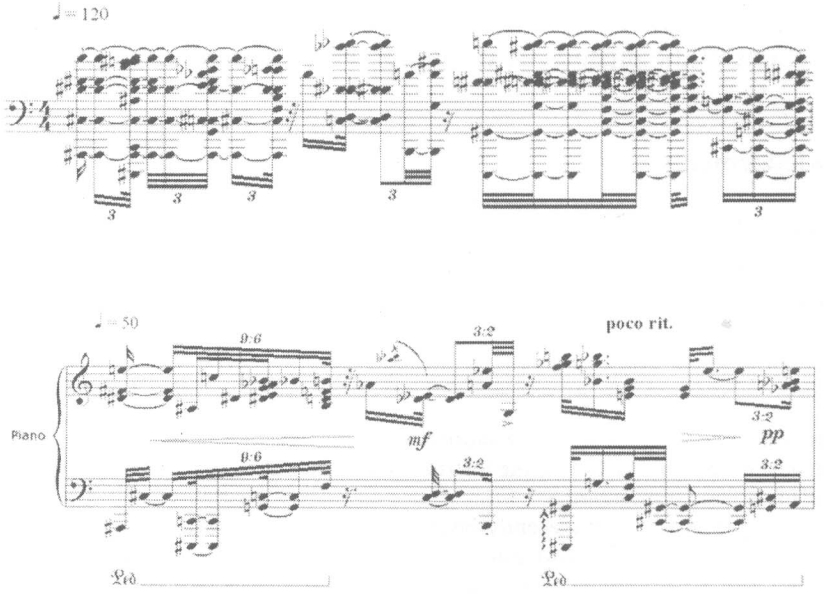

Figure 2. Example of instrumental writing (solo piano) derived directly from extracted data from speech analysis. By opening and converting the original MIDI file (saved from Open Music) in the notation software Sibelius, "messy" results are obtained (top line), which in turn can be treated as new raw material for further elaboration. The second line shows one possible result obtaind from it after compositional "sculpting".

Conclusions

The main question to a composer using the processes described above is: where and how to transform the raw data? Factors like intelligibility of words and of

other intermediate degrees of speech features are to be considered according to a given musical project. The use of the original sample in subtle mixings with the resulting synthesis has proved to be an efficient way to increase inteligibility withouth loosing the freshness of the new synthetic textures. So far we have used simple Csound orchestras files and sine waves as the timbre source for resynthesis. Naturally, further developments and different results can be obtained by designing more complex structures at those levels. At present, we are working on assembling the results of those first experiments on a multi-channel electro-acoustic piece[4].

The instrumental approach also showed promising results, although a higher level of composer's interference is needed when moving from a MIDI file to a more refined instrumental writing. Straight adaptations of converted data are often impossible. Note distribution among instruments and within a single instrument, rhythmic adjustments and orchestration issues are among the main factors to be considered in such adaptations. Finally, another step to be taken is to explore the possibilities of generating musical forms based on temporal and spectral manipulations of the prosodic structures obtained through linguistics analysis of speech.

Acknowledgements

Special thanks to composer Ignacio de Campos, who first taught me the basis of all the sound processing explained above, and to composers Larry Polansky and Eric Lyon for their comments and support while I was writing this paper.

Notes

[1] This has been a continuous source of musical ideas for new music composers, not only in the electro-acoustic domain, but also in the instrumental domain. Circles (1960), by Luciano Berio, is just one of many famous examples.

[2] Actually this problem was solved recently (beginning of 2004) with the use of a Perl script to sort out all the raw, non-averaged data in a suitable manner. Perl proved to be more efficient for this task than Open Music. This improvement was recently presented in a paper session during the 2004 Conference of the Society for Electro-Acoustic Music in the United States (SEAMUS), and will be published in the future.

[3] The amplitude in dB from Audiosculpt had to be rescaled to match the Csound dB amplitude scale.

[4] Most of the processes described here were also used in some of our recently composed pieces "Fonepoemas" (2003) and "Gedankenfabrik" (2003).

References

DODGE, Charles; JERSE, Thomas. *Computer Music*. New York: Schirmer, 1985.

HAYES, Bruce. *Metrical Stress Theory*. Chicago, University of Chicago Press, 1995.

JAKOBSON, Roman; HALLE, Morris. *Fundamentals of Language*. The Hague: Mouton & Co., 1956.

ROADS, Curtis; STRAWN, J. (organizers), *Foundations of Computer Music*. Cambridge: MIT Press, 1985.

ROADS, Curtis. *The Computer Music Tutorial*. Cambridge: MIT Press, 1996.

Bruno Ruviaro é compositor natural de São Paulo,SP. Formou-se em piano e composição pela Universidade Estadual de Campinas (2000) e atualmente está no segundo ano do curso de mestrado em música eletroacústica no Dartmouth College (EUA). Trabalha tanto com meios instrumentais como eletroacústicos, além de desenvolver pesquisa em improvisação de música contemporânea.