SIMULATION OF HUMAN VOICE TIMBRE BY ORCHESTRATION OF ACOUSTIC MUSIC INSTRUMENTS

Thomas A. Hummel
Experimentalstudio der
Heinrich-Strobel Stiftung des
SWR
Kartäuserstr.45
D79102 Freiburg
thomas.hummel@swr.de

ABSTRACT

This paper describes a method which simulates the timbre of the human voice through the orchestration of classic music instrumental ensembles. The spectral envelope of speech is used as a model for the orchestration. The idea of this method is to build sums of spectral envelopes of different sounds of music instruments in order to approach to the spectral envelope of a phoneme. Large instrumental databases with standardised intensities like the *Virtual Orchestra* are required. For each sound of the database, an averaged spectral envelope is calculated. Sounds with suitable spectral envelopes are chosen and get part of the orchestration. The resynthesis of phoneme sequences are considerably clearer than the resynthesis of a single phoneme. The simulation of a whispered word is better perceived as a voiced word.

Among other pieces, this method was applied in the orchestra piece *Nicanor*, premiered 1999 in Stuttgart. A considerable similarity of the respective orchestral sequence to the sound of the whispered word is perceived.

1. INTRODUCTION

1.1. Electroacoustic synthesis

The synthesis of the human voice, the artificial human voice, is one of the most important fields of interest of the computer music research, and on acoustic research in general. The aim is to create an illusion of a speaking or singing human voice, although nobody was speaking or singing. Since several decades research is evolving - with increasing success. Normal computer system software is now capable to speak text with significant quality.

Speech has a number of acoustic properties, such as pitch, rhythm and timbre. Phonemes, which are the atoms of speech, are acoustically defined as a timbre or a timbre transition. In all western languages, semantic content is exclusively imparted by the row up of phonemes.

Different models are the basis of synthesis approaches. One of the oldest models is the additive model, the synthesis of overtone series. This method is an empiric and unflexible approach, as it only allows the synthesis of vowels.

Another approach is the famous CHANT system [1]. It is using a physical model of the human voice and of music instruments in general, the fof method (formes d'ondes formantiques). It was developed in the 1980s at the IRCAM and yielded a powerful and realistic approach understanding the phonation process, but again was more suitable for vowel synthesis and the singing voice. In the 1990s, programs like the *Diphone* system by Xavier Rodet yielded a better again similarity to speech [2].

The common basis of all those approaches is the means of electroacoustic synthesis. The precision of computer synthesis promised a maximum of success.

1.2. Instrumental synthesis

On the other hand, composers interested more and more for the challenge of a purely instrumental music, which audibly creates the human voice and its timbre.

Empiric approaches to speech timbre have a long music history, like the work of Vinko Globokar [3] in the piece *toucher*. A more scientific approach may be found in pieces like *Im Januar am Nil* by Clarence Barlow [4] or *etymo* by Luca Francesconi [5]. It uses the idea of the orchestration of overtone series in order to create vowel-like timbres. Each instrument of a larger ensemble plays one partial of a large overtone series in its specific dynamic resulting in a vowel like orchestration sum (see Figure 1). Leaving aside strong compositional restrictions of this method, the reliability is limited by the fact that each involved instrumental sound has its own overtone series thus adulterating the result.

2. SUMS OF SPECTRAL ENVELOPES

2.1. Principle

In this publication, a new method is presented which synthesises speech-like timbres with music instruments. It is based on a model of speech as a sequence of spectral envelopes. This description is more general than the description of speech as a sequence of overtone series, as it also describes consonants or in general whispered speech. As soon as the sequence of spectral envelopes of a spoken sentence is determined, white noise may be filtered with these envelopes.

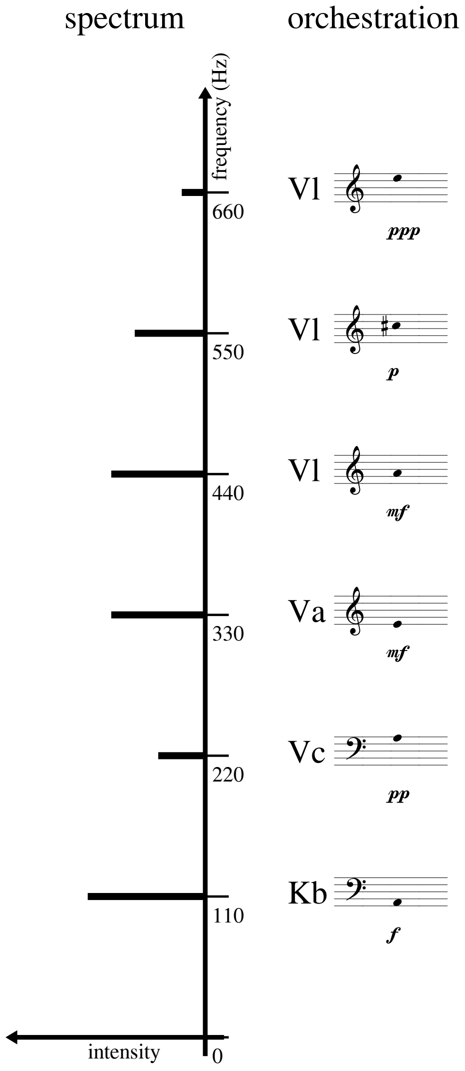

Figure 1 Orchestration of partials with a string ensemble.

The semantic content of the original text will be perceived from the resulting sound. This is the vocoder principle.

A sound of a music instrument also contains an evolution of spectral envelopes. If the sound is static, the spectral envelope is static too. Hence, if we consider static sounds, then we may sum up spectral envelopes of instrumental sounds to an overall envelope. The quantity to add is not the amplitude of the sound, as this would not consider the phase randomisation. Instead, the intensity of a sound, i.e. the squared amplitudes takes also phase effects into account (formula 1).

$$I = \sum_{i=1}^{n} A_i^2 \tag{1}$$

It should be possible to approach to the sound of a phoneme with a suitable set of instrumental sounds. The sequence of spectral envelopes of the speech sample corresponds to a sequence of orchestrations.

2.1. Calculation of spectral envelopes

Spectral envelopes are intensity averaged spectra. For the purpose of this investigation, a 2048 point FFT was calculated. A series of approximate minor third interval frequencies was defined as the base of the envelope: 55.00 Hz, 110.00 Hz, 185.00 Hz, 261.63 Hz, 311.13 Hz etc.

All FFT bands were assigned to the closest frequency band of the envelope. Intensities of the appropriate FFT bands were summed up to give the intensity of the envelope.

2.2. Sound databases

As the spectral envelope of different sounds of an instrument depend on many factors and may thus not be predicted easily, it is necessary to take advantage of instrumental sound databases. For this investigation, the *Virtual Orchestra*[1] was used. Such a database comprises many thousands of sounds from different instruments, different playing modes, different pitches and different dynamics. Unusual playing modes are as well considered as normal playing modes.

For each sound, a spectral envelope is calculated. For unstable sounds, the averaged envelope from five equidistant times within the sound is calculated. A large set of envelopes may be the base of an orchestration search.

2.3. Error minimisation

The principle of the orchestration search is the successive minimisation of the simulation error.

The first step approaching a suitable orchestration is simple. Within the database, the sound with the best resemblance of its envelope in relation to the original envelope is identified. We are thus minimizing the area difference Δ between the searched envelope and the envelope of the selected sound. It is important that the intensity of the selected envelope does not exceed in any frequency band the intensity of the origin, so that negative differences are avoided (exception see below). This is the first sound of the searched orchestration.

The residual error itself is an envelope. In a second approach, another sound is selected, which minimises this residual error, but again is at no frequency stronger than the residual error. This is the second sound of the searched orchestration (figure 2).

This process may be repeated until

- No more sound may be found to minimise the error
- The envelope is simulated sufficiently well
  (a given percentage of the area is covered)
- A given instrument ensemble plays tutti.

[1] The virtual orchestra is a commercial contemporary music sound and multimedia library including software. It was developed at the Experimentalstudio der Heinrich-Strobel-Stiftung des SWR

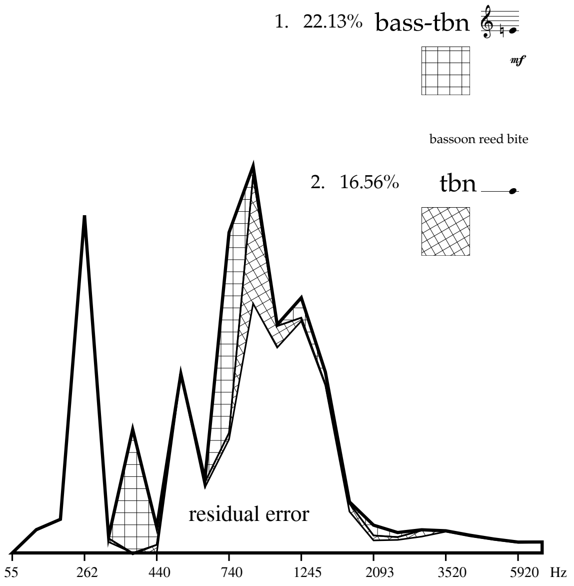

Figure 2 Successive minimisation of the residual error in a spectral orchestration

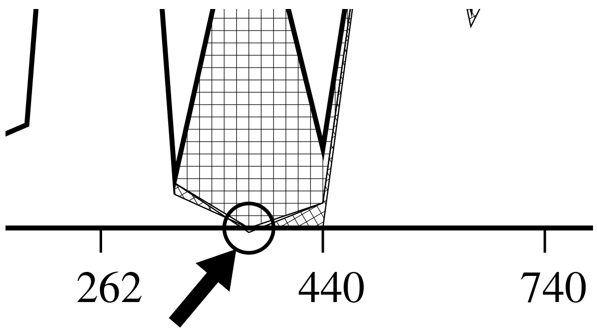

Figure 3 Tolerance in the minimal intensity of the residual error

Experiments with this optimization algorithm shows that the process is aborted soon when at some frequency of the envelope the residual error falls near to zero. In this case, only a few or even no instrumental sounds may be found which have a sufficiently weak intensity at this frequency. If a certain tolerance of exceeding intensity is allowed, the envelope sum is not significantly adulterated (figure 3). For the examples described here, the tolerance was set to 0.1% of the maximal intensity of the searched envelope. Figure 4a shows an orchestration of a female whispered „a", figure 4b the orchestration of a rolled „r".

3. PERCEPTABILITY

The experiments show that the resemblance of the original phoneme and its orchestration depends on several factors. A scrutiny of different phonemes like „a" and „r" reveal, that the resemblance of an - in terms of achieved percentage successful - optimisation is not necessarily well audible. The following factors have to be taken into account.

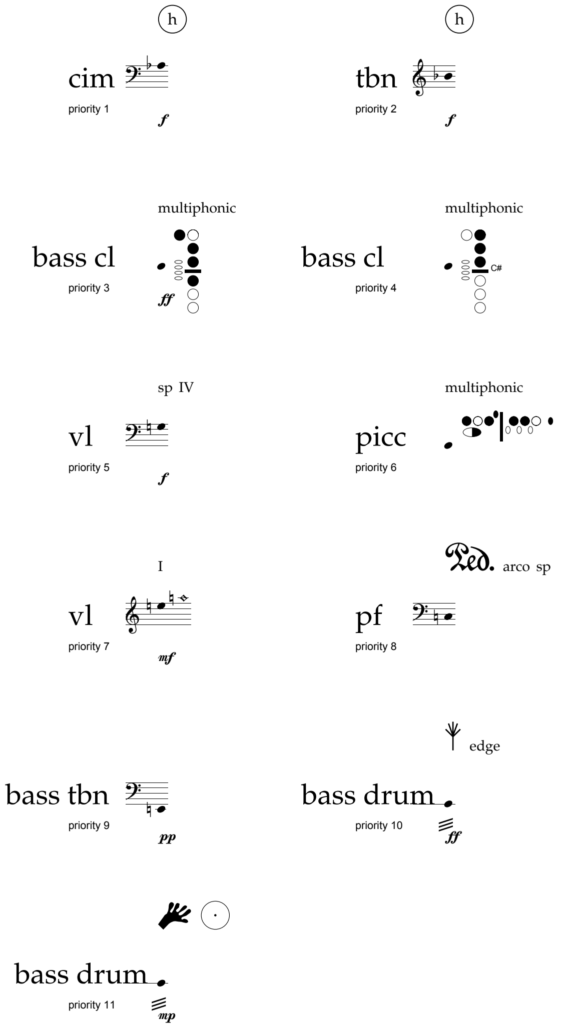

Figure 4a Orchestration of an "a" phoneme

3.1. Pitch perception

The pitch structure of the origin and the simulation is not correlated, as the algorithm only deals with frequency regions, not with frequencies. On the other hand, the human ear bases the similarity of sounds also on pitch similarity. In fact, the orchestration of a whispered sound has audibly a better similarity, as the whispered sound has no distinct pitches which may contradict to any pitches of the orchestration.

3.2. Sequencing

The power of the method is specially revealed, if a whole word is sequenced as a serie of orchestrations. It is well known that isolated or freezed phonemes are not as well recognised as a transitional sequence of phonemes within a word.

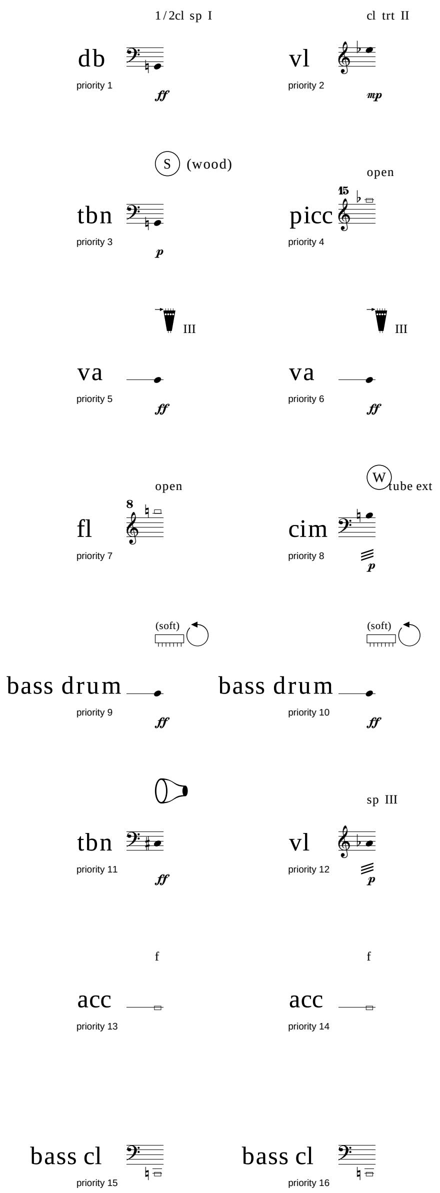

Figure 4b Orchestration of a rolled "r" phoneme

Same is true for the orchestrations. As an example, the spoken word *Nicanor* (meaning see below) was sliced into 20 equidistant time positions, and orchestrations were calculated for any time position. The assemblance of the orchestrations using samples gave a very clear perceptability of the word, although the speed of the orchestration transitions would be unplayable for an ensemble.

3.3. Harmonic/noise ratio

Fricative or plosive phonemes have high noise content. If voiced origins are used, it is useful to determine the noise content of the phoneme and to try to mirror this in the orchestration. The database of the *Virtual Orchestra* offers the noise content of each sound, which may be used as an additional condition for the choice of sounds.

4. USE IN COMPOSITION

The orchestration of phonemes is a compositional method, which may be generalised. Additional restrictions like registers, choice of instrument groups, noise amounts may be applied during the optimisation process and result in „closer" or less „close" results.

4.1. Nicanor

This method of speech simulation was first used by the author in the orchestral work *Nicanor* from 1996/1997, and since then in several other pieces (*bruillards* for string quartet, *Strietschech* for speaker and seven instruments, *From Trachila* for voices and orchestra, *Kopfwelten/Versteinerung* and, most recently, *Ins Ohr geschrieben*). *Nicanor* was premiered in the festival éclat Stuttgart/Germany in 1999. *Nicanor* is based on a novel by Garcia Marquez with the title *el otoño del patriarca*.

For this piece, the central passage of the novel reports from a fictive latin american dictator, who oppressed his people a whole life long and who murdered all his opposition. When getting old, he retires into his fortress. Nevertheless, the death passes all walls and calls him with the name *Nicanor*, as he calls all human beings in the moment of dying.

In so far, no man, but the hereafter is talking to the dictator. The function of the hereafter is taken by the orchestra. In the performance of the piece, the word *Nicanor* is quite well perceivable, although not as well as in the computer simulation - this is due to the inaccuracies of the interpretation.

5. CONCLUSION

The method of synthesis of phonemes by music instruments requires large digital sound databases, which are available since several years. It proposes detailed orchestrations, which sound through their similarity to phonemes equilibrated. On no account, the orchestrations follow classic orchestration rules.

6. REFERENCES

[1] Rodet, X. "The CHANT project", *Computer Music Journal 8(3*), MIT-Press, 1984.

[2] Rodet, X. et al. "Diphone sound synthesis based on spectral envelopes and harmonic/noise excitation functions", *Proceedings of the International Computer Music Conference*, Cologne, Germany, 1988.

[3] Globokar, V. *toucher* for a speaking percussionist, C.F. Peters, Frankfurt, Germany, 1973

[4] Barlow, C. *Im Januar am Nil* for ensemble, feedback edition, Cologne, Germany, 1982

[5] Francesconi, L. *etymo* for soprano, electronics and chamber orchestra, Ricordi 1994