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 instrumental ensembles for classical music. The spectral envelope of speech is used as a model for the orchestration. The idea of this method is to synthesise the spectral envelope of a phoneme using the spectral envelopes of different sounds of music instruments. Large instrumental databases with standardised intensities like the *Virtual Orchestra* are required. For each sound of the database, an average spectral envelope is calculated. An error minimisation algorithm optimises the similarity of the orchestration's envelope in relation to the phoneme. 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 is speaking or singing. Research has been evolving for several decades with increasing success. Normal computer 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 sequence of phonemes.

Different models are the basis of synthesis approaches. One of the oldest models is the additive model, the synthesis of overtone series. It only allows the synthesis of vowels.

Another approach is the famous CHANT system [1]. It uses a signal 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 to understand 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 again a better similarity to speech [2].

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

1.2. Instrumental synthesis

As an alternative technique, composers became more and more interested into the challenge of a purely instrumental music which aims to simulate 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]. 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 at its specific dynamic resulting in a vowel like orchestration sum. 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 compromising the result.

2. SUMS OF SPECTRAL ENVELOPES

2.1. Principle

In this publication, a new method is presented which synthesizes speech-like timbres with music instruments. It is based on a model of speech as a sequence of spectral envelopes. A spectral envelope of a sound is the boundary of its spectral properties. This model does not only describe pitched sounds (vowels), but also noisy phonemes and in general whispered speech.

Any sound with a spectral envelope evolution similar to speech resembles speech. This is the vocoder principle. Typically, white noise is filtered to obtain a sound with a certain spectral envelope.

Phases are not considered in this model. They are mainly important for plosives.

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 aim is to find a method which selects a set of sounds for which the overall envelope resembles the envelope of a phoneme.

2.1. Calculation of spectral envelopes

A spectral envelope s is the boundary of its spectral properties. It is here described as a series of discrete intensity /frequency band pairs:

$$s = i_1 f_1, i_2 f_2, i_3 f_3, , ... i_n f_n \quad (1)$$

The spectral envelope is calculated according to the procedure described in [5]. For the purpose of this investigation, a 2048 point FFT is calculated.
27 frequency baselines are defined at intervals equal to critical bandwidths according to [6]. For the most part of the spectrum, the interval is about a minor third:

110.00 Hz (a2), 185.00 Hz (f#3), 261.63 Hz (c4), 311.13 Hz (eb4), 369.99 Hz (f#4) …. 14080 Hz (a9).

FFT bins are assigned to the closest frequency band of the envelope. The intensities of all FFT bins related to an envelope band are summed up to give the intensity of a frequency band of the envelope.

2.2. Sums of spectral envelopes

As the phase correlation between different sounds is not defined, sums of intensities of frequency baselines vary with the time offsets of the involved sounds. Therefore, we calculate an average intensity sum for all frequency baselines. Intensities I are squared amplitudes A. For a given phase difference Δ, the intensity of the sum of two vectors a and b is calculated according to (2).

$$I_{sum} = A_{sum}^2 = A_a^2 + A_b^2 + 2A_a A_b \cos(\Delta) \quad (2)$$

The average over all phase differences Δ from 0 to 2π reveals that the averaged sum is just the sum of the squared amplitudes (3).

$$I_{sum,\ average} = A_a^2 + A_b^2 = I_a + I_b \quad (3)$$

Hence, we can just add up intensities of corresponding spectral envelope bands to get the averaged intensity sums.

2.3. 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.

For each sound, the absolute amplitude in dBA has been measured. Therefore, the volumes of all sounds are standardised. The spectral envelope is calculated for each sound and is standardised using the absolute amplitude. For unstable sounds, the averaged envelope from five equidistant times within the sound is calculated. A large set of envelopes is thus available as a base of an orchestration search.

2.4. Error minimisation

A spectral envelope is a multidimensional vector. A first approach to realise a spectral envelope vector as a sum of spectral envelopes is to derive an orthogonal vector system from the sound database. The sounds of the vector system are scaled to synthesise the wanted envelope vector. In a musical context, vector scaling would correspond to the dynamic of a sound. Unfortunately, the form of a spectral envelope of a sound from an instrument changes with the dynamic. Secondly, the sound is limited to a certain dynamic range. The sound vector thus has a nonlinear behaviour und is not suitable as part of a vector system.

Another approach seems to be more promising. This is the method of error minimisation. The first step approaching a suitable orchestration is simple. Within the database, the sound with the best resemblance of its envelope to the original envelope is identified. We are thus minimising the difference Δ between the target envelope S (capitals I,F) and the envelope of the selected sound s (small letters i,f).

$$\Delta = S - s = I_1F_1 - i_1f_1 ,\ I_2F_2 - i_2f_2 ,\ ...\ I_nF_n - i_nf_n \quad (4)$$

It is important that the intensity of the selected envelope does not exceed in any frequency band the intensity of the target, so that negative differences are avoided (exception see below). This is the first sound of the target orchestration.

The difference Δ is the residual error. The residual error itself is an envelope. In a second approach, another sound is selected, which fits best to the residual error, but again is at no frequency stronger than the residual error Δ. The residual error is thus minimised again. This is the second sound of the target orchestration.

[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 (description in http://www.ircam.fr)

step	Back-ground (figure 1)	Instrument (figure 2)	Enve-lope share	Residual error
1		Cimbasso	30.67%	69.33%
2		Eb-clarinet	21.60%	47.73%
3		Piccolo	16.76%	30.97%
4		Accordeon	8.41%	22.56%
5		Alto flute	7.04 %	15.52%
6		Bb-clarinet	5.20%	10.32%
7		Violin	2.98%	6.34%
8		Trombone	2.56%	3.78%
9		Double bass	1.25%	2.53%
10		Violin	0.96%	1.57%

Table 1. Minimisation of the residual error in a spectral orchestration (whispered "a")

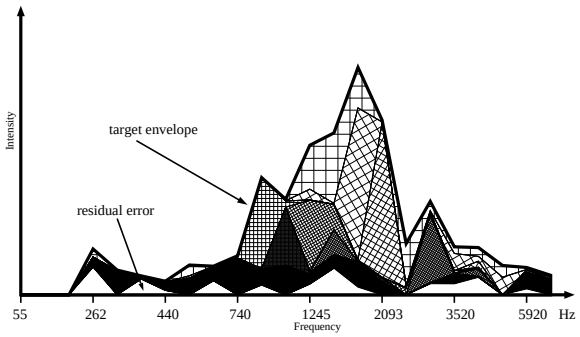

Figure 1 Minimisation of the residual error in a spectral orchestration (whispered "a")

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.

Table 1 shows a ten step error minimisation for a whispered "a". Envelope share means the area of the envelope in relation to the area of the target envelope in percent. Fig. 1 shows the corresponding spectral illustration. Fig. 2 shows the explicite orchestration in musical notation form. Experiments with this optimisation algorithm show that the process is aborted soon when the residual error in some frequency band of the envelope falls near 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 negative excess intensity is allowed, the envelope sum is not significantly compromised (Fig. 3). For the examples described here, the tolerance was set to 0.1% of the maximal intensity of the searched envelope.

The resulting orchestrations are musically extremely interesting as they are independent from any classical orchestration rules. They mix materials

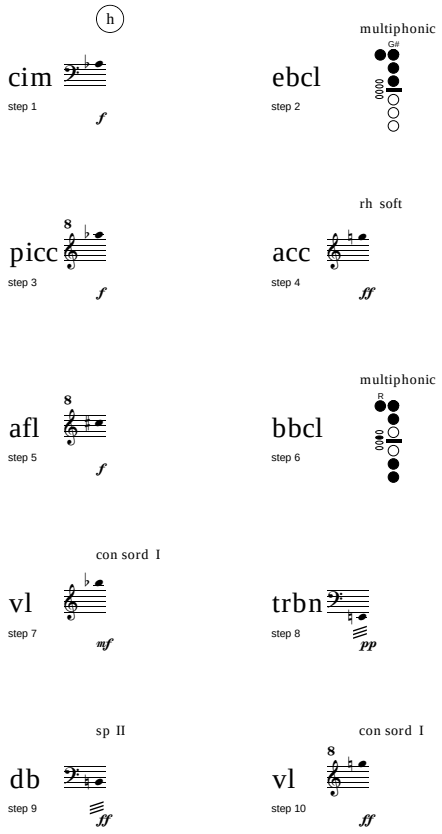

Figure 2 Orchestration of a whispered "a"

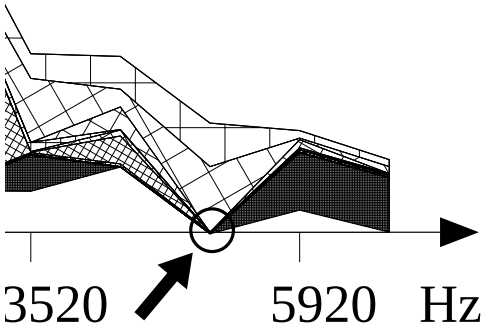

Figure 3 Negative tolerance in the minimal intensity of the residual error

together, which are completely heterogeneous in pitch, dynamic and playing technique. A huge labyrinth of surprising instrumental soundscapes is opened for the composer's activity.

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 a successful optimisation - in terms of achieved percentage - is not necessarily well audible. The following factors have to be taken into account.

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 judges the similarity of sounds also based 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 be in conflict with any pitches of the orchestration.

3.2. Sequencing

The power of the method is recognised especially when a whole word is sequenced as a series of orchestrations. It is well known that isolated or frozen phonemes are not as well recognised as a transitional sequence of phonemes within a word.

The same is true for the orchestrations. As an example, the spoken word *Nicanor* (meaning see below) was sliced into 20 equal time intervals, and orchestrations were calculated for any interval. The assemblage of the orchestrations using samples gave a very clear perceptibility 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 „realistic“ or less „realistic“ results.

The 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 the novel *el otoño del patriarca* by Garcia Marquez.

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. As he gets old, he retires into his fortress. Nevertheless, the death passes through all walls and calls him by the name *Nicanor*, just as he calls all human beings in the moment of dying.

Thus, no man, but the hereafter is talking to the dictator. The role 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.

Finally, a sound of the trumpet and a sound of the bass drum are simulated by the orchestra in some part of the piece.

5. CONCLUSION

The method of synthesis of phonemes by music instruments requires large digital sound databases, which have been available for several years. It proposes detailed orchestrations, which sound equilibrated through their similarity to phonemes. The method may be applied to other sound sources than the human voice. On no account, the orchestrations follow classic orchestration rules.

6. ACKNOWLEDGMENTS

The author thanks Diemo Schwarz for helpful comments.

7. 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] manual of the software/database *The Virtual Orchestra*, Appendix. Calculation of Acoustic information. Experimentalstudio/IRCAM 2003

[6] Zwicker, E. et al., “Critical bandwidth in loudness summation”, *Journal of the Acoustical Society of America* 1957