SIMULATION OF THE VIOLIN SECTION SOUND BASED ON THE ANALYSIS OF ORCHESTRA PERFORMANCE

Jukka Pätynen, Sakari Tervo, Tapio Lokki[*]

Aalto University School of Science
Department of Media Technology
Espoo, Finland
firstname.lastname@aalto.fi

ABSTRACT

A study on simulating a string instrument section of a symphony orchestra is presented. An anechoic recording of a single violin is used as the actual instrument sound, and the section sound is created by introducing temporal differences to the violin recording with a phase vocoder in the time-frequency domain. In addition, amplitude modulation and pitch shift are applied to the duplicated signals. The applied temporal differences follow a distribution that is obtained from a real violin section, recorded with contact microphones and analyzed with onset detection. Results of a listening test show that the proposed method provides a better simulation of a large instrument section than the traditional chorus effect.

Index Terms— strings section simulation, chorus effect

1. INTRODUCTION

The string section in a symphony orchestra produces a broader sound than a string instrument soloist or a chamber ensemble. Individual instruments in a section are not perceived separately, in contrary, their sounds blend together. A section sound results from the differences in playing technique, individual instruments, and the acoustic conditions. Meyer has stated that the sound characteristic for a section is caused by the broadening of the peaks at harmonic frequencies [1]. The intonation, i.e. nominal pitch of the played notes, is different. With instrumental ensembles the 3 dB bandwidth of the spectral peaks deviate up to ±20 cents from the nominal frequencies. Also, individual string instruments exhibit perceivable differences in their frequency response and resonance properties [2].

In audio industry the impression of a string section is commonly sought after. An audio effect, *chorus*, is used as an industry standard when a single instrument is needed to sound more like an ensemble [3]. The chorus effect is based on a delay line whose tap point is modulated over time, causing variation in tempo and pitch [4, 5]. Multiple delays can be applied for the corresponding number of simulated players.

Outside music industry, high-quality anechoic recordings are required in auralization. A small number of true anechoic orchestral recordings exist [6, 7], and the instruments are recorded one at a time due to practical reasons. Recording a large group of the same instruments would escalate the task considerably. However, for realistic orchestra auralization a plausible section sound is needed, thus various methods for ensemble simulation has been applied to anechoic recordings. Fixed prime delays up to 23 ms has been utilized in [8]. In [9], anechoic string recordings have been processed with a phase-synchronous overlap-add algorithm, but the varying acoustic reflections due to the different source positions have been observed to have a greater effect than the algorithm itself. The present authors have previously suggested fixed delays and pitch shifts for the simulation of the string section sound [10].

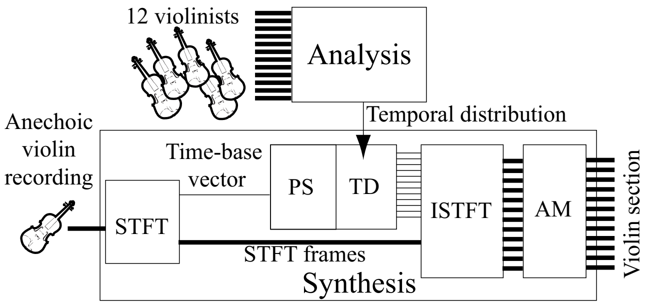


Figure 1: Block diagram of the proposed method. STFT, ISTFT, AM, PS, and TD stand for short-time Fourier transform, inverse short-time Fourier transform, amplitude modulation, pitch shift, and time difference, respectively.

In this paper, it is proposed that an instrument section can be simulated in a more natural manner if the temporal differences in an orchestra performance are taken into account. First, we extract data on the temporal differences between string players by detecting note onsets from contact microphone recordings. Then, the approximate temporal distribution is applied in a proposed simulation method for creating a string section sound from a single instrument recorded in an anechoic chamber. The principal motivation of this study is to improve the authenticity in auralization and loudspeaker orchestra - type applications [11]. The block diagram of the proposed approach is depicted in Fig. 1.

2. TEMPORAL ANALYSIS OF THE VIOLIN SECTION

The timing differences of the string players can be investigated with several approaches, such as accelerometers or IR tracking. In this article, the timing differences of the violinists are studied from the

[*] The research leading to these results has received funding from the Academy of Finland, project nos. [218238 and 140786], the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [203636], Finnish Foundation for Technology Promotion, Nokia Foundation, and Helsinki Doctoral Programme in Computer Science.

978-1-4577-0693-6/11/$26.00 ©2011 IEEE

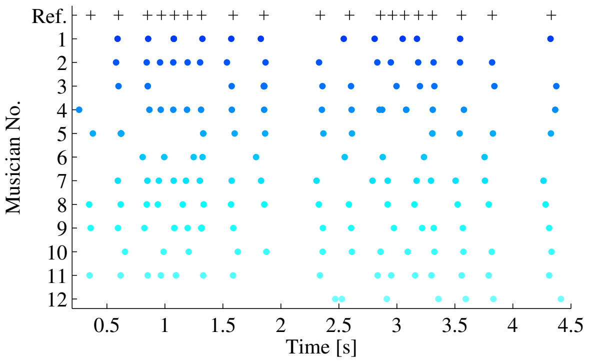


Figure 2: An example of the onset detection for $M = 12$ musicians. Bars 167-168 of Sibelius' Symphony no. 3, I movement, are shown.

note onsets. This is done by attaching contact microphones to the bridge of each violin in a section playing the same part and detecting the note onset from the captured signals. Contact microphones are beneficial for greatly reducing crosstalk between recorded channels.

Several well-developed methods exist for the onset detection [12]. Here, the spectral difference (SD) method is used due to its simplicity, and it is calculated as

$$f(n)^{(m)} = \sum_k [H(\|Y^{(m)}(n,k)\| - \|Y^{(m)}(n-1,k)\|)]^2 \qquad (1)$$

where $H(y) = (y + \|y\|)/2$, $Y(n,k)$ is the short-time Fourier transform of the signal at time index n and discrete frequency bin k, and m is the musician number. The onset detection is done in 85 ms time windows at every 0.5 ms. Additional smoothing with a 25 ms average filter for this function is applied, as suggested in [12]. The detected onsets are then the local maxima of the SD detection function $f(n)^{(m)}$.

A reference for the onsets is calculated from the combined SD detection function

$$F(n) = \Pi_{m=1}^{M} f(n)^{(m)} \qquad (2)$$

where M is the number of analyzed musicians. The combined SD detection function is filtered as previously the individual ones, and the local maxima of the function are selected as the reference.

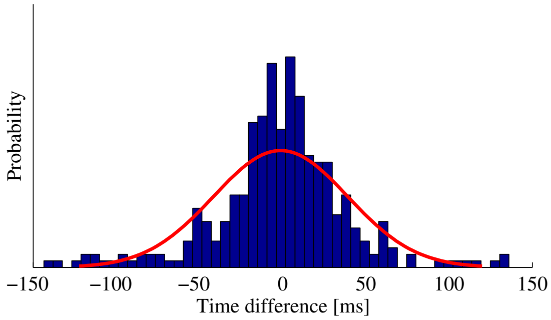
The detected onsets are grouped with respect to the reference. If an onset is within ±150 ms of the reference onset it is considered to belong to the same group. If multiple onsets exist for a single musician, the closest one to the reference is selected.

Each detected onset o_t at reference onset t is normalized with respect to the group normal with M musicians, i.e.,

$$\hat{o}_t^{(m)} = o_t^{(m)} - 1/M \sum_{m=1}^{M} o_t^{(m)} \quad . \qquad (3)$$

The final distribution is calculated over all the normalized groups. Missing data, i.e., missing onsets, are simply neglected from the results.

The onsets are calculated for a section of $M = 12$ violinists in an amateur orchestra. The selected passage is from Sibelius' Symphony No. 3, I movement, bars 167-181. An example of the onset detection results are shown in Fig. 2. In total, rates for false positive and false negative detections are 0% and 18.4%, respectively. Although the investigated orchestra was not professional, the chosen passage is relatively easy, and it was played with a good tone and intonation. Some results on the distribution of the onsets are shown in Fig. 3. Relying on the central limit theorem, the data can be expected to be normal, since the mean is 0.00 ms. To conclude, the temporal differences between the note onsets of the violin players are approximately normally distributed with 40 ms standard deviation in the studied passage.



σ	μ	2.5 %	25 %	50 %	75 %	97.5 %
40.35	0.00	-101.85	-18.34	0.00	19.10	84.20

Figure 3: Distribution of the onset temporal differences. Mean μ, standard deviation σ, and selected percentile statistics are given in milliseconds. Solid line shows a fitted normal distribution for comparison.

3. SIMULATION OF A STRING SECTION SOUND

The task is to simulate a string section from a single instrument recording, utilizing the results from the temporal analysis in Section 2. The proposed simulation method consists of three components (Fig. 1). First, a constant pitch shift is introduced for simulating intonation differences between individual musicians. Second, the timing of the playing is continuously varied to simulate the differences in note onsets between musicians based on the temporal analysis. Third, amplitude modulation is applied for simulating the differences in the velocity of the played notes. The proposed method is abbreviated in the following as PTA (Pitch-Time-Amplitude).

The signal recorded from a single instrument in an anechoic chamber is denoted by $x(t)$. Pitch shift and timing differences are implemented using the short-time Fourier transform (STFT) with a frame length of 2048 samples with 75% frame overlap. The signal in the discrete time-frequency domain is denoted by $X(n,k)$. The method is based on modifying the time-base vector of the STFT result with the phase vocoder [13]. In the following, the synthesis part in Fig. 1 is described.

3.1. Pitch shift

The constant pitch shift is implemented with an inverse time-stretch approach. A pitch shift for musician m is achieved by scaling the time-base vector for the changed playback rate by the factor of $d^{(m)} \in \mathbb{Q}$ closest to the desired detune. The playback rate is defined with a new time-base vector $\hat{n}^{(m)}$ that is the original vector n resampled by $d^{(m)}$ intervals. At this point the resampled time-base vector $\hat{n}^{(m)}$ causes a time-stretch in $X(n,k)$. The reciprocal pitch shift is subsequently obtained by resampling a time-domain signal by the inverse factor $1/d^{(m)}$, back to its original length.

3.2. Timing differences

The time-variance is implemented by adding random fluctuation to the new time-base vector $\hat{n}^{(m)}$. Here, a random Markov chain following the random walk Metropolis-Hastings sampling from normal distribution is used [14]. The advantage in using Metropolis-Hastings sampling is that the values in the random chain follow any given probability distribution after the burn-in sequence. With a low frequency, the random chain emulates the effect of a musician playing slightly behind the average rhythm instantaneously, and at the next moment catching up the tempo, or vice-versa. This is similar to the behavior of the traditional chorus effect but without its characteristic vibrato. Low-frequency random chain is interpolated in order to smooth the otherwise abrupt changes between the values in random sequence.

The time-base vector yielding a time stretch with varying time differences is obtained by combining $\hat{n}_{\mathrm{r}}^{(m)} = \hat{n}^{(m)} + r^{(m)}$, where $r^{(m)}$ denotes the interpolated random Markov chain. The time-frequency frames are sampled with the new time-base vector $X_m(\hat{n}_{\mathrm{r}}^{(m)}, k)$ using the phase vocoder approach, and then inverse-transformed back to the time domain. The desired pitch shift is finally achieved by resampling the time-domain signal by $1/d^{(m)}$.

3.3. Amplitude modulation

Amplitude modulation is applied to the time-domain signals in order to simulate the varying playing dynamics between musicians and between consecutive notes. Suitable modulation curves can be obtained similarly to the temporal variations, that is, with Metropolis-Hastings sampling. Low-frequency random sequence is generated, and the sum of parallel random values is scaled to unity. Hence, the amplitude modulation does not effect to the combined signal level, and only the balance between the simulated musicians is varied.

4. SUBJECTIVE EVALUATION

The authenticity of the proposed method for simulating the violin section sound was evaluated with a listening test. Eleven experienced subjects having their background on acoustics and/or signal processing participated in the test in which they rated the section sound in five processing conditions with two signals.

4.1. Stimulus signals

Two excerpts of anechoic violin recordings were utilized for comparing the processing methods. A professional violinist played short, six-second passages of Mahler's and Beethoven's symphonies representing typical orchestral repertoire (Mahler: 1st Symphony, IV movement, bars 57-61, II violin; Beethoven: 7th Symphony, I movement, bars 14-15, I violin). The violin was recorded with 22 calibrated microphones evenly positioned around the musician [6].

Five conditions were created from the anechoic signals. First, an unprocessed recording in one direction was taken to represent a solo violin performance as a reference. Second, 11 copies of the same recording was processed with individual chorus effects for creating an impression of a violin section. Third, the proposed PTA method was similarly applied to the original recording. Fourth, instead of one microphone signal, recordings from 12 different directions were processed with the identical PTA method. Fifth, the differences between unique violins were experimentally introduced with 11 filters whose magnitudes at the four Dünnwald bands were randomized between ±6 dB in order to alter the timbre of the violin [2]. After filtering, identical PTA processing was applied also here. In total, one solo performance and four violin sections of 12 simulated players were obtained.

The chorus effect for each copy was implemented as a linearly interpolated variable-length delay line without feedback, following the example given in [3]. Suitable parameters were chosen iteratively to produce a desired impression within the abilities of such a chorus effect. The randomized delay lengths for the individual copies were between 0-25 ms. Modulation signals were low-pass filtered white noise with the cutoff frequency at 3 Hz. Modulation depth was 1.3 ms. Lower values were considered to introduce too small differences and higher values for the modulation yielded unnaturally fuzzy results. These values fall within the guidelines in the literature [3, 4]. In PTA processing, the pitch shifts of the individual copies were distributed within ± 10 cents [10]. Temporal variation followed a normal distribution having a 45 ms standard deviation. For the normally distributed amplitude modulation the standard deviation was 1 dB with 5 Hz modulation frequency, which corresponds approximately to eighth notes in moderate tempo.

The processed dry signals were then convolved with room impulse responses measured in an unoccupied concert hall with a loudspeaker orchestra [11]. In the current study, three pairs of two Genelec 1029A -loudspeakers were positioned on the stage as the first violins. First of the two loudspeakers in each pair was positioned on a stand in 1 m height facing the stalls. The second one was on the floor facing upwards positioned 1 m further on the stage and 1 m to the side from the first loudspeaker. Directivity of the two-loudspeaker setup is closer to that of a violin than a single loudspeaker. Using a GRAS 3-D microphone probe, spatial impulse responses were recorded. For each source, the spatial response was rendered into two virtual cardioid microphones as a coincident XY pair with 90 degree separation. The processed copies of the anechoic recording were evenly assigned to the three sources and convolved with the corresponding impulse responses. The levels of the convolved signals were equalized with A-weighting.

4.2. Test setup

The subjects were asked to assess the perceived impression of a string section on a continuous linear scale. End points of the scale were ”one or few individual instruments” and ”large section with many instruments”. The subjects were instructed before the test that in an authentic section the individual instruments are not perceived as such. Instead, they are blended together, yet without artifacts or artificial coloration. The subjects were allowed to familiarize themselves with the signals and the test procedure before the test. The test for each condition and signal was repeated three times in a fully random order. The test was conducted in a quiet, acoustically treated listening room. The convolved stimuli were presented to subjects with Sennheiser HD650 headphones.

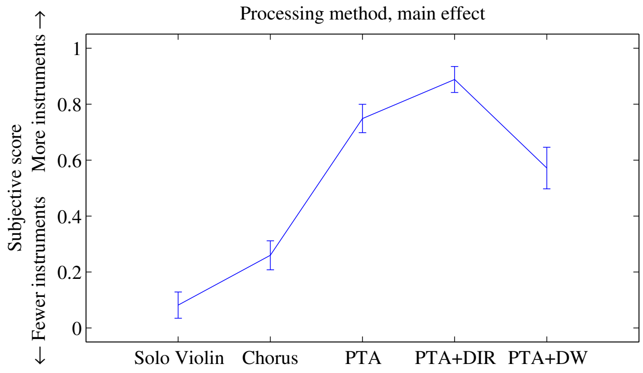


Figure 4: Results of the listening test shown with means and 95% confidence intervals. Higher score indicates larger perceived section.

4.3. Results

The five conditions were compared 66 times (11 subjects × 2 signals × 3 repeats). The results were analyzed with ANOVA having four factors: method, music, repeat, and subject. The results for the processing method are shown in Fig. 4. Higher location on the y-axis indicates a more convincing simulation of the section sound. The differences between all five conditions were significant ($F(4, 329) = 182.18, p = 0$). Solo violin condition received expectedly the lowest rating. All three variations of the proposed method were assessed to give an impression of a larger instrument section than the applied chorus effect. Utilizing different microphone directions (PTA+DIR) improved the impression compared to the PTA method applied to one microphone signal. However, altering the violin frequency responses with random Dünnwald band was not assessed better. No significant differences were observed between the subjects, repetitions, or signals.

5. CONCLUSIONS

The proposed PTA method, based on measured timing differences from a real orchestra, simulates an instrument section from a single recorded instrument. The results indicate that the method produces a better simulation for the section sound than the traditional chorus effect. The most prominent improvement is estimated to come from the large and varying time differences between the simulated instruments. The simulated section sound receives an improvement from using signals recorded from different directions.

Defining well-suitable attributes for the described listening test is challenging. A tentative listening test was organized previously, and the subjects were asked their subjective preference on a set of different processing methods and the principal attributes behind their judgement. The obtained attributes varied greatly along with the differences in the personal preferences. The evaluation of the attribute used in the present paper yielded much more reliable results.

While the PTA method can be implemented also with a series of delay-based effects, time-frequency processing conveniently allows the implementation of future extensions. These can be, e.g., altering tone transients [3], or modifying the frequency response based on detected note pitch [2]. With more extensive measurements, playing profiles could be created for different orchestras, particularly with the pitch estimation implemented.

6. REFERENCES

[1] J. Meyer, *Acoustics and the Performance of Music*. New York, NY, USA: Springer, 2009.

[2] C. Fritz, I. Cross, B. Moore, and J. Woodhouse, “Perceptual thrsholds for detecting modifications applied to the acoustical properties of a violin,” *J. Acoust. Soc. Am.*, vol. 122, no. 6, pp. 3640–3650, 2007.

[3] U. Zölzer, Ed., *DAFX:Digital Audio Effects*, 2nd ed. Chichester, United Kingdom: John Wiley & Sons, 2011.

[4] J. Dattorro, “Effects design, part 1: Reverberator and other filters,” *J. Audio Eng. Soc.*, vol. 45, no. 9, pp. 660–683, 1997.

[5] D. Kahlin and S. Ternstrom, “The chorus effect revisited-experiments in frequency-domain analysis and simulation of ensemble sounds,” in *Proc. 25th EUROMICRO*, Milan, Italy, Sep 8-10 1999, pp. 75–80.

[6] J. Pätynen, V. Pulkki, and T. Lokki, “Anechoic recording system for symphony orchestra,” *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 856–865, 2008.

[7] M. C. Vigeant, L. M. Wang, J. H. Rindel, C. L. Christensen, and A. C. Gade, “Multi-channel orchestral anechoic recordings for auralizations,” in *Proc. ISRA*, Melbourne, Austalia, Aug 29-31 2010.

[8] M. C. Vigeant, L. M. Wang, and J. H. Rindel, “Investigations of orchestra auralizations using the multi-channel multi-source auralization technique,” *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 866–882, 2008.

[9] T. Lokki, “How many point sources is needed to represent strings in auralization?” in *Proc. ISRA*, Seville, Spain, Sep. 10-12 2007, paper P11.

[10] T. Lokki and J. Pätynen, “Applying anechoic recordings in auralization,” in *The EAA Symposium on Auralization*, Espoo, Finland, Jun. 15-17 2009.

[11] J. Pätynen, S. Tervo, and T. Lokki, “A loudspeaker orchestra for concert hall studies,” in *The Seventh International Conference On Auditorium Acoustics*, Oslo, Norway, Oct. 3-5 2008, pp. 45–52, Also published in Acoustics Bulletin 2009, 34(6), pp. 32-37.

[12] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, “A tutorial on onset detection in music signals,” *IEEE Tr. Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[13] M. Dolson, “The phase vocoder: A tutorial,” *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[14] S. Chib and E. Greenberg, “Understanding the metropolis-hastings algorithm,” *American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.