Synthèse d'un jeu polyphonique à partir d'une seule voix - Rapport de mi-parcours

Maxime ARBISA, Prof. Dr. Stefan Weinzierl

Mars 2016 - Juillet 2016

Présentation du sujet

Mon stage se déroule du 1er Mars 2016 au 30 Juillet 2016, sous la direction du Prof. Dr. Stefan Weinzierl dans le laboratoire audio de l'Université Technique (TU) de Berlin. Le sujet s'articule en deux parties :
La première consiste à faire la synthèse d'un ensemble d'instruments à partir de l'enregistrement *anéchoïque* - c'est-à-dire un enregistrement du son pur, sans réverbération ni réflexion du son - d'un seul instrument. En d'autres termes, on doit être capable de simuler à partir d'un enregistrement anéchoïque de violon un ensemble de violons, comme dans un orchestre. Les enregistrements anéchoïques d'orchestre sont en effet rares voire inexistants car les studios qui les permettent sont souvent trop petits pour accueillir un grand nombre d'instrumentistes en même temps.
L'enregistrement anéchoïque présente ensuite l'avantage de pouvoir être placé dans n'importe quel environnement sonore. C'est cette deuxième partie qui m'a poussé à choisir le laboratoire audio de la TU de Berlin, puisqu'il est spécialisé dans la *synthèse binaurale dynamique*, et la simulation d'environnements acoustiques. Un de leur travail des plus remarquables consiste à recréer l'ambiance acoustique d'une salle de concert ou d'un auditorium disparu à partir des plans de l'édifice. On utilise ainsi un enregistrement anéchoïque qu'on fait résonner dans la salle de concert souhaitée. La synthèse binaurale vous permet ensuite de l'écouter comme si vous étiez vous-même assis au milieu de la salle.
Mon travail pendant ce stage permettra donc de recréer tout un orchestre à partir de l'enregistrement anéchoïque d'une seule trompette, d'un violon, d'un hautbois, en multipliant leur voix, et de le placer dans un environnement acoustique qui n'existe plus. La synthèse binaurale fait le reste : vous mettez votre casque sur les oreilles, vous êtes alors au troisième rang de l'opéra de Vienne, brûlé en 1785, en train d'écouter la 40ème symphonie de Mozart.

Le code, la bibliographie, et les résultats audio obtenus sont disponibles en temps réel, et en accès libre sur GitHub à l'adresse suivante :
`https://github.com/MaximeArbisa/AudioSynthesisFinal`

Explication du procédé

Lorsqu'un ensemble d'instruments joue, J. Meyer [8] remarque un élargissement des pics aux fréquences harmoniques ; la bande des -3dB grandit jusqu'à 20 cents autour des fréquences de l'enregistrement. J. Pätynen et al. [10] remarquent quant à eux que les attaques des violons d'un ensemble ne sont pas toutes jouées en même temps et suivent une loi normale $\mathcal{N}(0, \sigma^2)$ de variance $\sigma^2 = 45$ms centrée en chacun des temps, donc de moyenne 0.
L'effet d'une section réside donc dans des instruments de hauteur (*pitch*) différente et proche, et ne jouant pas tout à fait en même temps.
La technique appliquée usuellement dans l'industrie musicale est celle du *chorus* [6] qui consiste à superposer le même signal en lui appliquant un retard variable dans le temps. Ce retard variable va créer un *ré-échantillonnage* local (donc une hauteur différente), ce qui va donner l'impression d'instruments de hauteur différente jouant légèrement en décalé. Le problème de cette méthode est qu'elle génère des enregistrements souvent de basse qualité audio et comportant un vibrato caractéristique qui n'est pas acceptable pour des applications de simulations d'environnements acoustiques (en anglais, *auralization*).

Méthode utilisant le Vocoder de phase

La première méthode que nous utilisons est développée par J. Pätynen et al. [10] et s'appuie sur le *Vocodeur de phase* reposant sur la *Transformée de Fourier à Court Terme (TFCT)*. Elle consiste en une modification de la hauteur (Pitch) de ± 10 cents, de l'ajout de fluctuations temporelles (Time difference), puis d'une modulation d'amplitude (Amplitude) pour créer chaque instrument de la section à partir de l'enregistrement original. C'est la méthode PTA.
Pitch shifting : le changement de hauteur s'opère avec un ré-échantillonnage du morceau suivi d'un élargissement temporel (*time stretching*) réalisé avec le vocoder de phase (TFCT). La TFCT consiste à parcourir le signal en trames de longueur N_w, et espacées d'intervalles I donc situées tous les kI, qui garantit un recouvrement d'au moins 75% du signal. Sur chacune de ces trames, on effectue une *Fast Fourier Transform (FFT)* de N_{fft} points ou *canaux* afin d'obtenir le contenu spectral de la trame. On a ainsi une représentation spectrale du signal en fonction du temps. Pour réaliser un élargissement du morceau d'un facteur *stretch* sans en modifier la fréquence, on calcule la fréquence instantanée w_s entre deux trames consécutives pour chaque *canal* (grâce à la différence de phase instantanée). On n'a alors plus qu'à *dérouler la phase* : on place donc les trames sur les instants de synthèse kR (avec R = *stretch**I), et on ajoute la correction de phase $w_s.R$ correspondante. La fréquence instantanée étant préservée, seul le déroulement temporel du signal est altéré par cette opération.
Time difference : pour simuler les fluctuations temporelles entre les différents instruments d'une même section, J. Pätynen et al. [10] ont recours à l'algorithme de Metropolis-Hastings, puissante chaîne de Markov qui tire intelligemment des valeurs suivant la distri-

bution normale décrite tout à l'heure. La description de cet algorithme est expliquée par S. Chib et E. Greenberg dans [13]. Cette chaîne prend en compte les tirages précédents. Si le violoniste est en retard, il rattrapera le tempo au tour suivant et vice-versa. Ce méchanisme ressemble un peu à l'effet de *chorusing*, mais sans son vibrato caractéristique. Ainsi, à la place de prendre les trames tous les kI échantillons du signal ré-échantilloné, on les prendra tous les kI+$\delta(k)$, avec $\delta(k)$ la fluctuation donnée par Métropolis-Hastings, on calcule w_s entre les trames $kI + \delta(k)$ et $(k+1)I + \delta(k+1)$, puis on les place sur les mêmes marques de synthèse que précédemment k*R, R = *stretch*.I. En ajoutant une fluctuation, on a un intervalle de $I + \delta(k+1) - \delta(k)$. Les trames doivent cependant toujours se recouvrir suffisamment (hypothèse du vocoder de phase) : il faut donc faire un filtrage passe-bas pour que les fluctuations varient lentement d'une trame à l'autre.
Amplitude modulation : pour simuler les différences de jeu d'une note à l'autre pour un instrumentiste, on utilise encore une chaîne de Métropolis-Hastings, mais cette fois avec une distribution $\mathcal{N}(1, 0.1^2)$ (variation de 10% dans l'amplitude). On effectue encore un filtrage passe-bas de fréquence de coupure 5Hz (qui correspondent à 8 notes dans le tempo modéré).

La technique d'étirement temporel décrite plus haut ne fait attention qu'à la *cohérence horizontale* des phases (d'un seul canal au cours du temps). Il faut cependant aussi faire attention au rapport des différents canaux entre eux (*cohérence verticale*), au risque de voir apparaître des effets indésirables propres au vocoder de phase, tels que les effets de "*phasiness*" ou "*transient smearing*", qui engendrent des pertes de sons. Ceci est souvent du au fait que les hypothèses selon lesquelles le signal varie lentement et les trames se superposent assez ne sont pas respectées. J. Laroche et al. [7] ont démontré une technique de "*phase locking*", qui permet de regrouper les phases sur les pics d'énergie entre les canaux. Elle permet de travailler avec un recouvrement des trames moins limitant (50% au lieu de 75%), qui compense nos fluctuations temporelles. Pour les signaux variant rapidement (*transients*), des solutions développées par A. Röbel [12], [11] ont été proposées. Elles n'ont pas encore été abordées.

Méthode reposant sur des techniques temporelles

On ne rajoute cette fois des fluctuations temporelles que sur les onsets en allongeant ou en raccourcissant les notes. Il faut donc effectuer une détection d'onsets [1], ou puisqu'on n'a qu'un instrument soliste, on peut faire de la détection de hauteur (*pitch*) pour trouver les notes. On utilisera alors le *produit spectral* ou la *transformée à Q-constant* - plus adaptée à la musique, et avec une meilleure résolution dans les graves - et ses versions optimisées décrites dans [5], [2].
Enfin, on allongera ou raccourcira chaque note, soit en réutilisant le *vocoder de phase*, ou une technique temporelle, TD-PSOLA [9] qui duplique ou élimine les périodes d'un son, puis fait un "*overlap-add*" pour en modifier la longueur.

Autres méthodes

Les méthodes reposant sur la TFCT ou sur des méthodes temporelles telles que TD-PSOLA fonctionnent très bien avec des instruments tels que la trompette, le violion ou le hautbois car leurs harmoniques sont claires et détectables avec une simple FFT. Ces méthodes rencontrent cependant des limites lorsqu'appliquées à des instruments de percussion, où la résolution de la FFT ne permet pas de détecter leurs fréquences caractéristiques très proches.
Les méthodes à Haute Résolution (méthodes HR) vues en cours seront plus appropriées pour ces cas-là. Je n'en ai parlé que brièvement avec mon superviseur, puisque nous sommes encore sur le violon, mais elles peuvent être des outils appréciés pour traiter les instruments restant de l'orchestre qui ne peuvent être étudiés avec les méthodes décrites plus haut.

Simulation dans un environnement acoustique

Une fois qu'on aura traité tous les enregistrements anéchoïques des instruments dont on aura multiplié les voix, et qu'on obtiendra l'enregistrement anéchoïque d'un orchestre d'une bonne qualité sonore, on pourra le placer dans l'environnement acoustique souhaité. Nous n'en avons pas parlé avec mon maître de stage, la première partie étant encore en cours de traitement mais la TU possède un logiciel qui permet de simuler n'importe quelle salle de concert, de placer les différentes sources sonores sur la scène, ainsi que la place d'écoute de l'orchestre. Ce logiciel provient d'études menées par S. Weinzierl et C. Büttner sur la simulation de salles de représentations disparues, et décrites dans [14], [3] et [4].
Ils ont aussi une salle entourée de hauts parleurs afin de recréer un espace acoustique immersif. On pourra alors y brancher les sorties du logiciel afin de créer une *synthèse binaurale*, et d'entendre l'orchestre généré comme si nous étions dans la salle.

Commentaires

Le stage se déroule plutôt bien pour le moment, avec des résultats satisfaisants. J'avance à une bonne allure, bien que laissé totalement indépendant. Les réunions avec mon encadrant sont rares pour le moment, mais je pense que ça se débloquera une fois que j'aurais fait le tour de plusieurs méthodes et que j'aurais des résultats qui peuvent être traités pour la synthèse binaurale, où il sera plus présent.
Ce n'est pas grave pour le moment, j'ai plein d'idées, et j'ai de bons rapports, assez fréquents, avec un doctorant qui travaille sur les mêmes sujets que moi.

Bibliography

[1] "A tutorial on onset detection in music signals". In : *IEEE Tr. Speech and Audio Processing* 13.5 (2005), p. 1035–1047.

[2] Judith C. Brown et Miller S. Puckette. "An efficient algorithm for the calculation of a constant Q transform". In : *The Journal of the Acoustical Society of America* 92.5 (nov. 1992).

[3] Clemens Büttner et Stefan Weinzierl. "The acoustics of early concert venues in Japan". In : *DAGA* (2010).

[4] Clemens Büttner et al. "Acoustical characteristics of preserved wooden style Kabuki theaters in Japan". In : (sept. 2014).

[5] Hans Fugal. "Optimizing the Constant-Q Transform in Octave". In : *Linux Audio Conference*. Avr. 2009.

[6] D. Kahlin et S. Ternstrom. *The chorus effect revisited-experiments in frequency-domain analysis and simulation of ensemble sounds"*. Milan, Italy, sept. 1999, p. 75–80.

[7] Jean Laroche et Mark Dolson. "Improved Phase Vocoder, Time-Scale Modification of Audio". In : 7.3 (mai 1999), p. 323–332.

[8] J. Meyer. "Acoustics and the Performance of Music". In : (2009).

[9] E. Moulines et J. Laroche. *Non parametric techniques for pitch-scale and time-scale modification of speech*. T. 16. Fév. 1995, p. 175–205.

[10] Jukka Pätynen, Sakari Tervo et Tapio Lokki. "Simulation of the violin section sound based on the analysis of orchestra performance". In : *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (oct. 2011).

[11] Alex Röbel. In : ().

[12] Alex Röbel. In : *Proc. of the 6th Int. Conference on Digital Audio Effect (DAFx-03)* (sept. 2003).

[13] E. Greenberg S. Chib. "Understanding the Metropolis Hastings algorithm". In : *American Statistician* 49.4 (1995), p. 327–335.

[14] Stefan Weinzierl et al. "The Acoustics of Renaissance Theatres in Italy". In : 101 (2015), p. 632–641.