CONTRÔLE GESTUEL DE LA SYNTHÈSE CONCATÉNATIVE EN TEMPS RÉEL DANS LUNA PARK

Georges Aperghis
Composer
georges.aperghis@wanadoo.fr

Grégory Beller
RIM
beller@ircam.fr

ABSTRACT

Dans cet article, sont présentés les recherches et les développements réalisés pour la création de Georges Aperghis, intitulée «Luna Park». Cette oeuvre est largement reliée, à différents niveaux, au paradigme de synthèse concaténative, tant envers sa forme que dans les procédés qu'elle emploie. Grâce à la programmation temps réel, des moteurs de synthèse concaténative et de transformation prosodique sont manipulés, contrôlés et déclenchés par le geste, via des accéléromètres réalisés pour la création. La création mondiale de «Luna Park» a lieu à Paris, dans l'espace de projection de l'IRCAM, le 10 juin 2011, dans le cadre du festival AGORA 2011.

1. INTRODUCTION

«Luna Park» est une oeuvre de théâtre musical, d'une durée d'environ une heure, écrite par Georges Aperghis, mise en scène par Daniel Levy, et dont l'informatique musicale est réalisée par Grégory Beller. Le thème général de la pièce aborde la façon dont la télé-surveillance et le recueil massif de données numériques personnelles font de notre monde d'aujourd'hui, un gigantesque parc d'attraction, d'où le titre de l'oeuvre. Les quatre interprètes sont Eva Furrer, flûte Octabasse et voix, Johanne Saunier, danse et voix, Mike Schmidt, flûte basse et voix, et Richard Dubelsky, percussions aériennes et voix. Le fait qu'ils aient tous les quatre des parties vocales à interpréter, ainsi que la scénographie de l'ensemble (comprenant vidéo et décors), rapprochent «Luna Park» d'une précédente oeuvre de G. Aperghis, appelée «Machinations» (2002). Toutefois, cette nouvelle oeuvre se distingue de la précédente, notamment par l'utilisation de capteurs de geste et de la synthèse vocale.

En effet, différents capteurs, (accéléromètres, rubans tactiles et capteurs piézo-électriques) ont été développés et réalisés pour permettre aux interprètes de contrôler différents moteurs audio, par le geste. Le mapping entre les données issues de ces capteurs et les différents traitements audio, réalisé au sein de l'environnement de programmation temps réel Max/MSP, est différent d'une séquence à l'autre et peu évoluer dans le temps. C'est pourquoi nous présentons cet article selon le plan suivant.

Dans une première partie, cet article recense les différents capteurs réalisés pour cette création et donne des détails sur leurs développements, ainsi que sur les données qu'ils produisent. Dans une seconde partie, les moteurs audio réalisés sont décrits sous la forme de processus temps réel. Outre le moteur de synthèse concaténative, est présenté un moteur de transformation prosodique innovant permettant la modification du débit de parole en temps réel. La troisième partie propose quelques exemples de «mapping» entre les capteurs et les moteurs audio, notamment utilisés pour l'oeuvre. Enfin, une quatrième partie permet de conclure et de proposer quelques perspectives.

2. CAPTATION DU GESTE

2.1.Background

Dans le cas des instruments acoustiques, le geste est nécessaire à la production sonore. Ceci n'est plus le cas des instruments électroniques où le contrôle de processus sonores électroniques est dissocié du processus de production sonore. Plusieurs projets de recherche et de création à l'IRCAM emploient la captation du geste. Que ce soit pour augmenter des instruments acoustiques, projet du violon augmenté (Bogen Lied [20], augmented violon Project [8]) ou projet de percussions augmentées de Fedele, ou bien pour diverses interactions musique-geste-danse liées à la création (Glossopoeïa) ou à la pédagogie [17], un certain nombre de capteurs ont été réalisés à l'IRCAM et interfacés avec des processus sonores et/ou vidéos. Toutefois, la captation du geste n'a pas encore été utilisée pour contrôler des modifications prosodiques et/ou des synthétiseurs vocaux dans le spectre de la création contemporaine à l'IRCAM. C'est donc une nouvelle manière d'utiliser les systèmes de captation du geste que nous proposons pour ce projet de recherche et de création.

Le contrôle gestuel de la synthèse vocale constitue désormais un champ de recherche à part entière. Différents types de contrôleur ont été élaborés pour différents types de synthétiseurs de voix parlée, ou de voix chantée [13]. Parmi ces mappings, on trouve le «speech conductor» [1, 2], le «Glove-Talk» [16], le «squeeze vox» [11], le «SPASM» [12], le projet «OUISPER» [18, 14], et d'autres [15]. Nous avons choisi d'utiliser les mouvements de la main pour plusieurs raisons, outre passant les raisons scénographiques. Tout d'abord, la parole spontanée peut être naturellement accompagnée d'un mouvement des mains. L'idée d'accompagner les mouvements des mains par de la parole, par réversibilité, semble donc naturelle.

L'aspect percussif des mouvements fait écho à la synthèse concaténative dans laquelle des segments sont déclenchés de manière discrète dans le temps, gérant ainsi l'aspect segmental de la parole. A l'inverse, l'aspect continu des mouvements des mains permet un contrôle de la prosodie, aspect supra-segmental de la parole. Si l'on considère la dissymétrie classique droite-gauche telle que la connaissent les chefs d'orchestre (pour les droitiers, la main gauche est plutôt reliée à l'expression, tandis que la main droite est plutôt reliée aux marqueurs temporels importants de la musique), on peut alors créer un contrôle gestuel des deux mains de la synthèse, avec pour un droitier, une main droite gérant l'aspect segmental et une main gauche gérant l'aspect supra-segmental. C'est un des scénarios possibles que nous avons exploité pour la création de la pièce (voir section 4).

2.2.Gants accéléromètres

La technologie des gants accéléromètres/gyroscopes sans fil utilisée [17] permet de mesurer les accélérations des deux mains selon 6 axes (3 en translation et 3 en rotation avec les gyroscopes). Les données brutes délivrées par les gants ne sont pas forcément aisées à interpréter, aussi une première étape de pré-traitement/mapping permet de rendre ces données plus interprétables.

2.2.1.Pré-traitements

Les données provenant du récepteur wifi sont transmises via udp toutes les 1 ms. Afin des les synchronisées au timing interne de Max/MSP, elle sont tout d'abord filtrées median (5) et sous-échantillonnées d'un facteur 5, c'est à dire que l'on obtient un flux stable de données toutes les 5 ms. Puis différents descripteurs du geste sont issus de ces données brutes.

2.2.2.Variation de la quantité de mouvement

L'estimation de l'accélération instantanée permet de connaître, à tout moment, la variation de quantité de mouvement relative au geste. Cette quantité de mouvement, selon les lois de la mécanique classique, et directement proportionnelle à la vitesse. Les données brutes venant du capteur sont d'abord «dé-bruitées» grâce à l'emploi d'une moyenne courante reposant sur les 4 derniers échantillons. La racine de la somme des carrées de ces six valeurs filtrées permet d'obtenir une grandeur proportionnelle à la variation de quantité de mouvement du geste.

2.2.3.Hit Energy Estimation

Le Hit Energy Estimation permet le déclenchement instantané à partir de l'observation de la variation de la quantité de mouvement du geste. Les trois valeurs, délivrées par les capteurs d'accélération en translation, sont stockées dans un buffer circulaire comprenant à tout instant, 20 échantillons. Les trois déviations standards correspondantes à ces valeurs sont, à tout instant, sommées (norme I correspondant à la somme des valeurs absolues). Cette somme permet aussi de représenter la variation de quantité de mouvement du geste. Afin de détecter des variation rapides de cette valeur, correspondant à des variations brusques du geste, elle est comparée à tout instant à sa valeur médiane (5). Lorsque la différence entre ces deux valeurs dépassent un certain seuil arbitraire, une valeur discrète apparait pour signifier la présence d'un changement rapide du geste. Cela permet, par exemple, d'émettre un click régulier, lorsque l'on bat une mesure avec la main de bas en haut, à chaque fois que la main change de sens.

Le Hit Energy Estimation est un processus permettant de générer des données discrètes à partir d'un geste, par définition continu. En effet, d'un signal physique continu, il permet par seuillage, de définir des instants correspondants aux pics de variation de la quantité de mouvement, qui coïncident, d'un point de vue perceptif pour l'utilisateur, à des pics d'efforts (d'accélération). Par ce procédé, il devient alors possible de créer des percussions aériennes précises ou les sons sont déclenchés au moment ou la main de l'utilisateur change de sens ou accélère subrepticement.

2.2.4.Orientation de la main

Il est assez difficile, voire impossible, d'obtenir un contrôle lent, quasi-statique ou de définir une position absolue dans l'espace, à partir d'accéléromètres et de gyroscopes. Toutefois, le champ gravitationnel terrestre introduit un offset dans la réponse des capteurs qui peut être exploité pour déduire l'orientation absolue de la main, ainsi que la présence de mouvements lents. Cette mesure quasi-statique apporte un contrôleur continu à l'interprète. Un exemple ludique de l'utilisation de ce type de donnée est le potentiomètre rotatif aérien grâce auquel la rotation de la main peut contrôler le volume (ou autre) d'un son (voir section 4).

2.3.Capteurs piézo-électriques

Outre les percussions aériennes, réalisables grâce aux gants accéléromètres, le percussionniste joue des percussions corporelles, c'est à dire qu'il émet des sons en frappant certaines zones de son corps et/ou de la structure l'entourant. L'intégration de zones locales et sensibles au toucher, sur un vêtement, n'est pas aisée. Les traditionnels «pads» de batterie sont trop rigides et non adapté à des frappes de main. Nous avons choisi d'utiliser des microphones piézo-électriques, plus petits et plus souples d'utilisation. Deux micros (un près de la hanche gauche et l'autre près de l'épaule droite) placé sur le percussionniste lui permettent de jouer avec deux zones distinctes de sont corps. Six micros du même type sont aussi disposés dans son espace environnant. Les signaux audio délivrés par ces différents micros sont traités par à un système de détection d'attaque classique permettant au percussionniste de déclencher différents types de sons selon les zones ou il frappe. Un peu comme si les «pads» d'une batterie

électronique étaient disposés sur et autour du percussionniste.

3. MOTEURS AUDIO

3.1.Synthèse concaténative

La synthèse concaténative est réalisée en temps réel grâce à l'objet Mubu.concat~ développé en collaboration avec l'équipe Interaction Musicale Temps Réel de l'IRCAM [21, 22]. L'objet prend, en entrée, un fichier son (buffer) et un fichier de marqueurs associé. Il permet la lecture des segments par le choix de leurs index. Il comprend aussi d'autres options affectant la lecture, comme la transposition, le fenêtrage, ou encore la sortie utilisée. Les segments peuvent se succéder automatiquement ou être déclenché par un métronome ou encore un signal discret émis par un capteur (issu du Hit Energy Estimation, par exemple). L'ordre des segments est arbitraire et toutes les séquences d'index sont possibles. Parmi elles, une série incrémentale de pas 1 restituera le fichier audio d'origine sans présence audible de la segmentation, tandis qu'une série aléatoire générera une variation du matériau de départ et rendra audible la segmentation préalablement choisie. Nous allons voir différentes méthodes pour générer des séquences d'index intéressantes dans le cadre de la synthèse de parole.

3.1.1.Synthèse de la parole

Si le moteur audio de synthèse concaténative n'a pas besoin d'être sophistiqué, comparé à d'autres paradigmes de synthèse, telle que la synthèse HTS ou encore la synthèse articulatoire, c'est parce que l'intelligence de la synthèse concaténative de parole repose sur la sélection des segments, c'est à dire, sur la définition de la séquence des index choisis. En effet, la synthèse concaténative de la parole repose sur deux distances permettant de définir simultanément la proximité du résultat par rapport à une cible (distance à la cible) et la qualité de ce résultat (distance de concaténation).

3.1.2.Distance à la cible

La première distance, comme son nom l'indique, nécessite la définition d'une cible. Dans la synthèse Text-To-Speech, cette cible est définie de manière symbolique par le texte et par les différentes analyses qui en découlent (grammaire, phonétique, ...). Dans la synthèse hybride parole/musique [7], une cible peut être définie de manière acoustique comme une séquence de descripteurs audio. A peu près n'importe quelle cible peut-être utilisée tant qu'elle partage avec les segments, un descripteur commun. Le synthétiseur IrcamTTS [4, 25], présente une particularité vis à vis des autres synthétiseurs TTS, puisqu'il permet à l'utilisateur de définir sa cible de manière symbolique, par le texte, mais aussi de manière acoustique, par des symboles prosodiques. Ainsi l'utilisateur peut écrire sur le même support, de manière conjointe, le texte souhaité et la façon dont il aimerait que ce texte soit prononcé. Cet outil est, par conséquent, très apprécié des compositeurs qui peuvent écrire non seulement le texte, mais aussi la prosodie qu'ils souhaitent, comme une partition. La cible peut aussi être définie en temps réel.

3.1.3.Distance de concaténation

La distance de concaténation permet d'évaluer le poids perceptif de la concaténation de deux segments. Des segments naturellement consécutifs occasionnent un poids nul. Des segments dont les spectres aux bords sont très différents, engendreront, quant à eux, une distance élevée, censée traduire la génération d'un artefact de synthèse.

3.1.4.Synthèse de la parole en temps différé

Comme la définition d'une cible en temps réel n'est pas chose courante, la plupart des synthétiseurs TTS fonctionnent en temps différé. L'utilisateur écrit une phrase, puis choisit généralement une voix, pour la synthétiser. L'algorithme de sélection des segments le plus répandu fait alors appel à un décodage Viterbi permettant la minimisation conjointe de la distance à la cible et de la distance de concaténation sur toute la phrase à synthétiser. La phase «backward» de l'algorithme permettant la définition de la solution optimale nécessite de connaître la fin de la phrase pour en synthétiser le début, ce qui rend le moteur de synthèse profondément non temps réel.

3.1.5.Synthèse de la parole en temps réel

La synthèse TTS de la parole en temps réel n'a de sens que si le texte est généré en temps réel. Cela peut-être le cas comme nous le verrons plus tard, grâce à des modèles statistiques (HMM, N-gram, K-NN ...) qui transforme ou génère une texte en temps réel. La contrainte «temps réel» ne nous permet plus d'utiliser la phase «backward» de l'algorithme de Viterbi garantissant l'optimalité du chemin sur toute une phrase, puisque nous ne connaissons pas à l'avance, la fin de la phrase en cours. La minimisation conjointe de la distance à la cible et de la distance de concaténation ne peut alors se faire qu'entre chaque segment, de manière locale et non globale. L'avantage de cette méthode réside dans sa réactivité, alors que son inconvénient est qu'elle produit un résultat sonore plus pauvre que la synthèse classique en temps différé.

3.2. Génération de la cible en temps réel

Dans «Luna Park», plusieurs paradigmes sont utilisés pour générer des cibles qui guident la synthèse concaténative en temps réel.

3.2.1. Séquences prédéterminées

Tout d'abord, certains textes sont fixés, à priori, et modélisés sous la forme d'une séquence de mots, de syllabes, de phones ou de semiphones. Les segments constituant ces séquences sont alors déclenchées par les

interprètes et peuvent être choisis en fonction de données issues des capteurs ou des analyses vocales (voir section Mapping). Cela permet d'avoir des suite de syllabes prévues, et de percevoir un sens sémantique clair.

3.2.2. Séquences prédéterminées et N-gram

Il arrive que plusieurs segments possèdent le même symbole (plusieurs phones correspondant au même phonème, par exemple). On peut alors évaluer les probabilités de transition d'un symbole à un autre et effectuer des tirages aléatoires respectant cette probabilité de transition. Grâce à un N-gram d'ordre N variable (plus N est grand, plus le tirage correspond à la séquence de départ et plus N est petit, plus le tirage est aléatoire), on peut contrôler en temps réel le rapprochement ou l'éloignement du matériau généré par rapport au texte prédéterminé au départ. Cela permet notamment aux interprètes de contrôler l'aspect sémantique de la synthèse en temps réel.

3.2.3. Ensembles prédéterminés et HMM

Ensuite, des ensembles de segments (syllabes, phonèmes, mots...) ont aussi été définis à priori. De la même manière, des segments appartenant à ces ensembles sont déclenchés par les interprètes et choisis selon des ensembles de descripteurs ou de manière aléatoire. Cela permet notamment de créer des textures (de phones) ou encore des rythmes à partir d'une seule syllabe. Une interface a été crée pour permettre de choisir et de faire évoluer en temps réel, la probabilité d'apparition de tel ou tel symbole. Elle se présente sous la forme d'un histogramme des symboles disponibles. L'édition de cet histogramme permet de modifier la probabilité de transition d'un symbole à un autre (HMM d'ordre 1). Une fois le symbole choisi, le segment déclenché peut être défini de manière aléatoire ou par différents contrôleurs/descripteurs.
La figure 1 présente l'interface qui correspond au texte «livre_P52bis_compte» dit par Richard Dubelsky le percussionniste. L'enregistrement a été segmenté automatiquement grâce au programme IrcamAlign [19] qui permet la segmentation de parole en phrases, groupes de souffle, mots, syllabes, phones et semiphones (en temps différé). L'unité de segmentation choisie, sur la figure 1, est la syllabe. L'histogramme en dessous de la forme d'onde a été mesuré sur le fichier entier (178 syllabes) et présente le taux d'apparition relatif de chacune de leurs symboles. L'édition de cet histogramme permet de modifier la probabilité d'apparition des syllabes. Par exemple, le second histogramme (tout en bas) provoque un tirage équiprobable de syllabes «ni», «naR» et «noR».

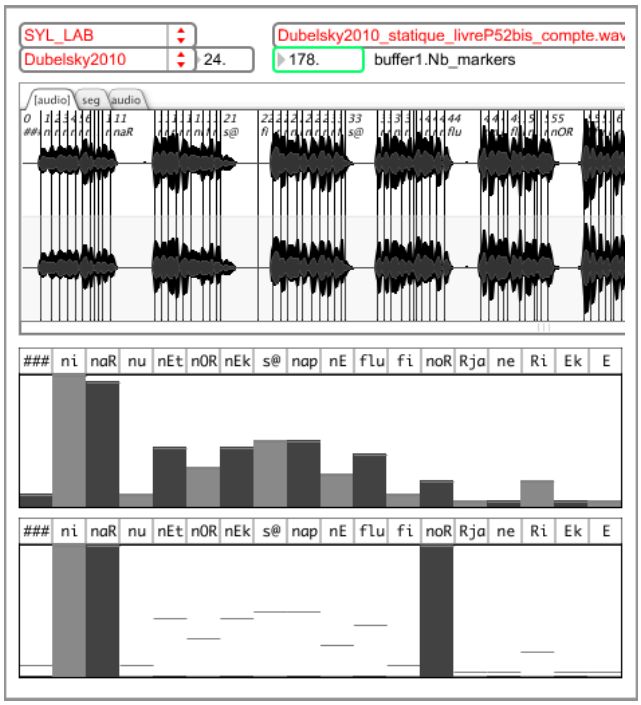

Figure 1. Interface de définition de la cible en temps réel par histogramme. En haut, un enregistrement segmenté en syllabes. Au milieu, l'histogramme présentant le taux d'apparition relatif de leurs symboles. En bas, le même histogramme édité permet de modifier la probabilité d'apparition des syllabes. Par exemple, un tirage équiprobable de syllabes «ni», «naR» et «noR».

3.3.Transformations prosodiques en temps réel

La modification de la prosodie en temps réel permet beaucoup d'applications [5]. Elle est possible grâce à un paradigme d'analyse-resynthèse de la parole permettant d'estimer et de modifier les dimensions prosodiques de la parole. Les transformations prosodiques peuvent s'appliquer à la voix des interprètes ou à la synthèse de manière équivalente.

3.3.1.Analyse prosodique en temps réel

La prosodie ou la manière de parler peut être décrite par un espace à cinq dimensions: L'intonation, l'intensité, le débit de parole, la qualité vocale et le degré d'articulation [3]. Si la plupart de ces dimensions sont aujourd'hui mesurables en temps différé [3], certaines le sont aussi en temps réel. C'est le cas de l'intonation (yin~ [10]), de l'intensité (loudness~). Nous avons ajouté un estimateur du débit de parole (syllex~). C'est trois descripteurs temps réel de la parole permettent de nous renseigner sur la façon dont parlent les interprètes. Ils peuvent être utilisés, au même titre que les capteurs, pour contrôler les différents moteurs audio.

3.3.2.Modification prosodique en temps réel

Si la modification de l'intonation par transposition et celle de l'intensité par gain variable sont désormais connues et assez bien maitrisées en temps réel, il n'en va pas de même pour le débit de parole. Or nous exposons dans cet article, une nouveau paradigme permettant la modification du débit de parole en temps réel. Toutes les transformations prosodiques utilisées reposent sur le moteur audio du Super Vocoder de Phase, SuperVP~ [25]. La librairie temps réel SuperVP~ permet déjà une transposition de qualité, ainsi que d'autres modifications comme celle de l'enveloppe spectrale. L'un des objets de cette librairie, SuperVP.ring~, permet d'effectuer ces transformations sur un buffer circulaire dont la taille peut être définie arbitrairement. L'utilité du buffer circulaire est de préserver l'instantanéité de la transformation, tout en autorisant, à tout moment, de pouvoir se déplacer dans le passé à court terme (terme équivalent à la taille du buffer). Grâce à cela, on peut localement allonger certaines portions du signal comme les voyelles (grâce à une détection de voisement en amont) et provoquer chez l'auditeur la perception d'un ralentissement du débit de parole. Si l'on ne peut avancer dans le future, le retour à la position instantanée peut se faire de manière accélérée, provoquant, cette fois, la perception d'une accélération du débit de parole. Comme si la tête de lecture du buffer se conduisait comme un élastique que l'on tend et détend. A l'extrême, il est possible de «geler» la position de lecture du buffer à un endroit, ce qui provoque un «arrêt sur son» qui peut donner des effets intéressants (tenue d'une voyelle par exemple).

4. EXEMPLES DE MAPPING

Dans cette partie, nous donnons quelques exemples de mapping entre les données de contrôle et les paramètres des moteurs audio. La figure 2 recense les différents contrôleurs disponibles (à gauche), ainsi que les différents paramètres des moteurs audio (à droite). Le mapping consiste à relier les contrôleurs aux paramètres (moyennant quelques transformation d'échelle linéaire ou non) et il peut varier, comme c'est le cas dans «Luna Park». Deux types de liaisons sont possibles: Les liaisons discrètes et les liaisons continues. En effet, les contrôleurs discrets (soulignés sur la figure 2), ne donnant qu'une valeur de temps en temps, comme le Hit énergie, correspondent à du contrôle de type percussif et vont servir à contrôler les paramètres discrets des moteurs audio, comme le déclenchement d'un segment pour la synthèse concaténative. A l'inverse, une liaison continue relie un contrôleur continu, comme la position linéaire absolue sur un ruban tactile à un paramètre continue des moteurs audio comme la transposition de la transformation prosodique. Nous donnons à présent à titre d'exemple, quelques scénarios choisis pour «Luna Park».

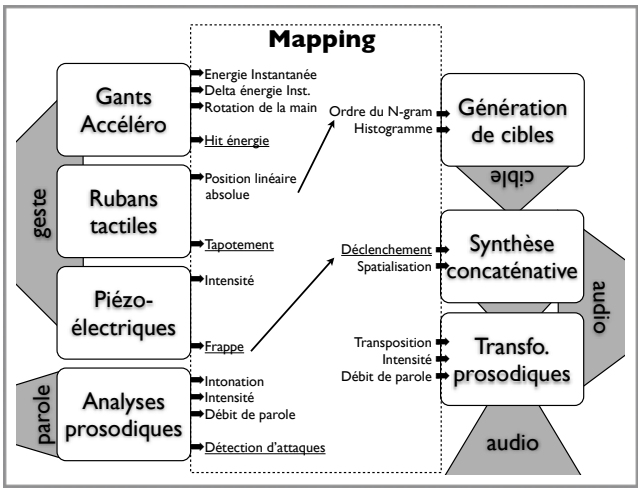

Figure 2. Le mapping est à l'interface entre les données issues des contrôleurs (à gauche) et les paramètres des moteurs audio (à droite). A titre d'exemple, deux flèches ont été tracées. La plus haute permet de faire varier l'ordre du N-gram par ruban tactile. La plus basse permet de déclencher un segment de la synthèse concaténative par une frappe sur le corps.

4.1.1.Prosodie aérienne

En reliant de manière directe le *Hit énergie* du gant droit du percussionniste (qui est droitier) au déclenchement de la synthèse, celui-ci peut gérer le débit de parole par des mouvements percussifs de la main droite. Si la *rotation de la main* du gant gauche est, quant à elle, reliée à la transposition et à l'intensité de la synthèse, il peut alors contrôler la prosodie de celle-ci avec les deux mains.

4.1.2.Je tiens ta langue

Dans une scène de «Luna Park», la danseuse Johanne Saunier, caresse du doigt un ruban tactile situé au-dessus d'un écran transversale. La vitesse de sa caresse, déduite de la position linéaire absolue détectée, est alors reliée à la modification du débit de la parole de Mike Schmidt, qui est en train de parler au même moment. Grâce à un mapping favorisant la transformation des voyelles, elle peut arriver à faire «un arrêt sur son» de la voix de Mike Schmidt, qui donne l'effet qu'elle lui tire la langue.

4.1.3.Il faut être sage pour comprendre

Le cumul de l'énergie instantanée des mains du percussionniste est utilisé, dans un scénario, pour contrôler l'ordre du N-gram de la génération de cibles. Le déclenchement de la synthèse est alors automatique (flux ininterrompu où chaque fin de segment, ici des phones, en déclenche un autre) et ne change que l'ordre dans laquelle les segments sont lus. Plus le percussionniste fournit de l'énergie et plus l'ordre du N-gram diminue, allant jusqu'à l'aléatoire pour des mouvements gesticulants de grande ampleur.

5. FUTURES PERSPECTIVES

La forme en tableau et l'attrait pour le phonème, qui sont deux constantes du compositeur G. Aperghis, ont permis d'explorer la synthèse concaténative, en temps différé et en temps réel, de manière assez complète. Son intérêt pour le théâtre musical a permis l'élaboration de scénarios reliant le geste à la vocalité. Une exploration exhaustive de ce rapport serait intéressante à mener car la parole est aussi un geste vocal. La comparaison des deux, leurs interactions ainsi que leurs sémantiques communes sont autant de terreaux fertiles à la création et à la recherche.

De cette période de recherche peuvent être déduites plusieurs perspectives concernant la synthèse vocale en temps réel, la captation du geste ainsi que leurs liaisons. Tout d'abord, il serait intéressant de se pencher sur d'autres synthétiseurs de parole basés sur des modèles paramétriques, semi-paramétriques ou hybrides. En effet, la synthèse concaténative en temps différé a pour avantage son degré de réalisme, qui devient difficile à maintenir en temps réel. Dès lors, elle devient malheureusement un synthétiseur difficile à paramétrer et perd de son intérêt en temps réel. D'autres modèles permettraient d'offrir une plus grande souplesse comme les modèles articulatoires, les modèles basés sur des HMMs ou encore des modèles hybrides concaténatif/HMM, qui sont la tendance actuelle.

Concernant la captation du geste, les capteurs utilisés possèdent l'avantage d'être assez légers pour être intégrer à des gants, ainsi qu'une bonne sensibilité au mouvement (captation de la variation de la quantité de mouvement). En revanche, ils présentent une autonomie énergétique assez faible (chose contraignante en production), et n'offre pas la mesure de la position statique absolue. Le rapport signal/bruit des capteurs ne permettant pas l'intégration des valeurs données pour en déduire une position statique, il serait bénéfique d'adjoindre aux accéléromètres, une autre technologie permettant d'accéder à la position absolue.

Enfin le «mapping» entre le geste et la synthèse de parole est un riche sujet de recherche, en plein essor, comme le montre notre participation au premier «International Workshop on Performative Speech and Singing Synthesis». Comme piste de recherche, on peut imaginer des «mappings» plus complexes où s'entremêlent les temporalités et les sémantiques du geste et du geste vocal.

6. ACKOWLEDGEMENT

Fred Bevilacqua, Bruno Zamborlin, Norbert Schnell, Diemo Schwarz, Riccardo Borghesi, Emmanuel Flety, Maxime Le Saux Pascal Bondu, Xavier Rodet, Christophe Veaux, Pierre Lanchantin et Jonathan Chronic.

7. REFERENCES

[1] C. d'Alessandro, N. D'Alessandro, S. Le Beux, J. Simko, F. Çetin and H. Pirker. «The speech conductor : gestural control of speech synthesis». In *eNTERFACE 2005*, The SIMILAR NoE Summer Workshop on Multimodal Interfaces, Mons, Belgium, 2005

[2] N. D'Alessandro, C. d'Alessandro, S. Le Beux and B. Doval. «Real-time CALM Synthesizer: New Approaches in Hands-Controlled Voice Synthesis.». In *Proceedings of NIME 2006*, pp. 266–71, 2006

[3] G. Beller. «Analyse et Modèle génératif de l'expressivié: Application à la parole et à l'interprétation musicale». In *PhD thesis, Université Paris XI, IRCAM*, June 2009.

[4] G. Beller, C. Veaux, G. Degottex, N. Obin, P. Lanchantin, and X. Rodet. «Ircam corpus tools: Système de gestion de corpus de parole». In *TAL*, 2009.

[5] G. Beller, «Transformation of expressivity in speech.» In *Linguistic Insights*, 97:259–284, 2009.

[6] G. Beller, D. Schwarz, T. Hueber, and X. Rodet. «Speech rates in French expressive speech». In *Speech Prosody 2006*, SproSig, ISCA, pages 672–675, Dresden, 2006.

[7] G. Beller, D. Schwarz, T. Hueber, and X. Rodet. «Hybrid concatenative synthesis in the intersection of speech and music». In *JIM*, volume 12, pages 41–45, 2005.

[8] F. Bevilacqua, N. Rasamimanana, E. Fléty, S. Lemouton, F. Baschet «The augmented violin project: research, composition and performance report». In *6th International Conference on New Interfaces for Musical Expression (NIME 06)*, Paris, 2006

[9] F. Bevilacqua, F. Guédy, N. Schnell, E. Fléty, N. Leroy, «Wireless sensor interface and gesture-follower for music pedagogy». In *Proc. of the International Conference of New Interfaces for Musical Expression (NIME 07)*, p 124-129.

[10] A. de Cheveigné and H. Kawahara, «YIN, a fundamental frequency estimator for speech and music». In *JASA,* 2002.

[11] P. Cook and C. Leider, «Squeeze Vox: A New Controller for Vocal Synthesis Models». In *International Computer Music Conference* (ICMC), 2000.

[12] P. Cook, «SPASM: a Real-Time Vocal Tract Physical Model Editor/Controller and Singer: The Companion Software Synthesis System». In Computer Music Journal 17/1 (1992), pp. 30–34.

[13] P. Cook, «Real-Time Performance Controllers for Synthesized Singing». In *NIME Conference*, 236_237, Vancouver, Canada.

[14] B. Denby and M. Stone, «Speech Synthesis from Real Time Ultrasound Images of the Tongue». In *IEEE International Conference on Acoustics, Speech,*

and Signal Processing, pp. I685-I688, Montreal, Canada, 2004.

[15] A. Esposito, M. Faundez-Zanuy, E. Keller, M. Marinaro, B. Kröger and P. Birkholz, «A Gesture-Based Concept for Speech Movement Control in Articulatory Speech Synthesis». In *Verbal and Nonverbal Communication Behaviours*, Springer Berlin / Heidelberg, 4775 174--189 (2007)

[16] S. Fels, & G. Hinton. «Glove-Talk II: A Neural Network Interface which Maps Gestures to Parallel Formant Speech Synthesizer Controls». In *IEEE Transactions on Neural Networks*, 9 (1), 205_212

[17] E. Fléty, C. Maestracci, «Latency improvement in sensor wireless transmission using IEEE 802.15.4», in *NIME 2011*, 2011

[18] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, M. Stone. «Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips». In Interspeech, pp. 2028-2031, Brisbane, Australia, 2008

[19] P. Lanchantin, A. C. Morris, X. Rodet, and C. Veaux. «Automatic phoneme segmentation with relaxed textual constraints». In *LREC2008*, Marrakech, Morocco, 2008.

[20] S. Lemouton, «Utilisation musicale de dispositifs de captation du mouvement de l'archet dans quelques oeuvres récentes». In *JIM 2009*, Grenoble 2009.

[21] N. Schnell, R. Borghesi, D. Schwarz, F. Bevilacqua, and R. Müller. «FTM—Complex Data Structures for Max». In *ICMC*, Barcelona, Spain, Sept. 2005.

[22] N. Schnell, A. Röbel, D. Schwarz, G. Peeters, R. Borghesi, «MuBu & Friends - Assembling Tools for Content Based Real-Time Interactive Audio Processing in Max/MSP». In *International Computer Music Conference (ICMC)*, Montreal, August 2009.

[23] D. Schwarz, G. Beller, B. Verbrugghe, and S. Britton., «Real-time corpus-based concatenative synthesis with catart». In *DAFx*, 2006.

[24] A. Roebel, F. Villavicencio and Rodet. «On Cepstral and All-Pole based Spectral Envelope Modeling with unknown Model order». In *Pattern Recognition Letters*. vol. 28, n° 11, p. 1343-1350, Août 2007

[25] C. Veaux, G. Beller, and X. Rodet. «Ircamcorpustools: an extensible platform for spoken corpora exploitation». In *LREC2008*, Marrakech, Morocco, may 2008.