Proposition de recherche doctorale

Synthèse et transformation expressive de la voix chantée

Mots clés :

- **Directeur de thèse** : Thomas Hélie
- **Co-encadrant(s)** :
- **Unité de recherche** : Sciences et Technologies de la Musique et du Son
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

Application d'algorithmes récents pour l'analyse et la charactérisation de la voix, pour la modélisation du chant et utilisation des modèles obtenus dans le contexte de synthèse et transformation du chant. {{Introduction}} La synthèse du chant est un sujet de la recherche sur la voix qui n'est que relativement peu traité dans la littérature. Les systèmes proposés couvrent la synthèse par FOF [Rodet 1984, Bennet 1989], la synthèse par concaténation d'unités [Macon 1997a/1997b, Bonada 2003, Kenmochi 2007], la synthèse par conversion de parole [Saitou 2007, Roebel 2012], et la synthèse par modèle HMM.. La plupart de ces systèmes ne prennent en compte que une petite part des connaissances sur le chant [Sundberg 87, Carlsson 1992]. Le système de synthèse par FOF développé à l'IRCAM dans les années 1980-1990 était une exception car il contenait un grand nombre de règles concernant les propriétés et caractéristiques du chant. Les algorithmes pour la transformation d'identité de la voix (parlée et chantée) sont devenus très performants et permettent d'obtenir des transformations avec une qualité très souvent tres proche des sons naturels [Banno 2007, Kawahara 2008, Degottex 2010, Roebel 2010, Roebel 2012]. Les algorithmes pour la caractérisation et la modélisation de la source glottique sont devenu suffisamment robustes pour être utilisés dans le contexte de l'analyse et de la synthèse de la voix chantée. Une application de ces algorithmes ouvre des perspective novatrices pour la transformation et la synthèse du chant [Roebel 2012]. Dans ce contexte nous proposons un travail sur l'application des algorithmes avancés d'analyse et de transformation dans le contexte de la synthèse et de la transformation de la voix chantée. Au départ des travaux, un modèle de contrôle pour le chant doit être établi qui permettra la description des paramètres d'un signal chanté à partir d'une partition. Un système pour la synthèse concaténative utilisant une base de signaux chantées sera mis en oeuvre pour effectuer la synthèse a partir d'une partition et d'un texte. Par la suite, doit être mis au point un algorithme pour l'apprentissage des paramètres relevants du modèle de contrôle, à partir d'exemples des signaux chantés, ceci afin de permettre une synthèse de chant dans le style d'un chanteur spécifique. {{Contexte local}} L'équipe « Analyse et Synthèse de Sons » de l'IRCAM poursuit des recherches portant sur l'analyse des signaux audio à des fins de transformation, de synthèse, de transcription, et d'indexation. L'analyse et la transformation de la voix sont des sujets centraux étudiés dans l'équipe. Plusieurs algorithmes pour l'analyse [Roebel 2007, Degottex 2010, Hézard] et la transformation de la voix [Peeters 2001, Degottex 2010, Roebel 2010, Roebel 2012] ont été développés dans l'équipe. D'autre sujets de recherche liés aux travaux à effectuer dans cette thèse, sont la conversion de la voix [Villavicencio 2010, Lanchantin 2010] et la synthèse concaténative de la parole à partir du texte [Veaux 2008]. Les travaux effectués dans cette thèse seront d'abord basés sur les algorithmes existants. Pendant la thèse les algorithmes seront adaptés en fonction des besoins identifiés.

Résumé du projet de recherche (Langue 2)

Les travaux à entreprendre pendant cette thèse sont découpés en trois phases: 1) Modèle de contrôle Afin de pouvoir établir un système de synthèse du chant, un certain nombre de paramètres décrivant le signale à générer (évolution du pitch, du vibrato, de l'énergie, etc) doivent être établis. La description de ces contours des paramètres du chant doit être suffisamment flexible pour pouvoir exprimer autant que possible toutes les stratégies d'un chanteur professionnel. Un jeu de paramètres permettant le contrôle de ces contours (par ex. fréquence du vibrato) doit être développé. La base pour ce contrôle sont les règles établies dans la littérature [Sundberg 1987, Bennet 1989, Carlsson 1992, Saino 2007] qui doivent être complétées notamment en y ajoutant des stratégies de contrôle des paramètres du pulse glottique. Pour cela une petite base de données du chant doit être établie et analysée afin de pouvoir étudier les contours des paramètres du pulse glottique. Ces analyses vont produire des connaissances nouvelles pour la description du chant. 2) Synthèse concaténative Le système de synthèse sera basé sur le principe de la synthèse concaténative. Par contre le système doit inclure des algorithmes de transformation afin de pouvoir effectuer les différentes ornementations (vibrato, crescendo, ..) qui sont utilisées dans le chant et afin de réduire la taille de la base de chant qui sera nécessaire pour synthétiser toutes les notes et tous les mots possibles. Beaucoup des algorithmes nécessaires sont déjà opérationnels avec une qualité suffisante. Par contre la manipulation des paramètres du pulse glottique, qui sera liée par exemple aux changements de type crescendo, demande de nouvelles approches de transformation basées sur la transformation du pulse glottique. La création de chant expressif, par exemple de style rock-and-roll, demande des transformations nouvelles qui rendent la voix plus rauque. Le déplacement des formants synchrone avec la hauteur des notes demandera une nouvelle approche pour la détermination des positions des formants. Les progrès scientifiques obtenus dans cette phase seront une base pour d'autres développements sur la transformation expressive de la parole et la conversion du locuteur. 3) Apprentissage du style de chant Les paramètres des contours de contrôle de la synthèse utilisés dans le point 1) permettent de modifier le chant synthétisé. Ces paramètres sont caractéristiques pour un style de chant et peuvent alors être appris à partir d'exemples. Dans cette dernière phase des travaux, un système pour l'apprentissage des paramètres doit être développé. Le système d'apprentissage recevra un certain nombre d'informations sur le contexte musical et adaptera les paramètre de contrôle en fonction des exemples du chant du chanteur cible. Cet apprentissage automatique des paramètres de contrôle semble essentiel pour une synthèse expressive du chant. Cela permettra une évaluation des stratégie de contrôle établie dans le point 1) et du système de synthèse développé dans le point 2) sur la base d'une comparaison entre le résultat du système et le chant du chanteur réel.

Informations complémentaires (Langue 2)

{{Perspective à longue terme}} Les algorithmes développés seront expérimentés dans le contexte artistique de l'IRCAM. Sur le long terme ils seront mis à disposition des artistes et compositeurs à l'IRCAM par le moyen d'une intégration dans les environnements pour la composition par ordinateur (OpenMusic) et pour le traitement du signal en temps réel (Max/MSP). Une perspective intéressante sera de connecter le moteur de synthèse chant temps réel avec l'outil pour le suivi de partition anticipative de l'IRCAM (Antescofo) afin de permettre une synthèse du chant adapté au contexte d'un concert "live". {{Bibliographie}} [Rodet 1984] X. Rodet, Y. Potard, J.-B. Barriere; "The CHANT project: from synthesis of the singing voice to synthesis in general". Computer Music Journal, vol 8, no. 3 pp. 15–31. [Sundberg 1987] J. Sundberg;"The science of the singing voice", DeKalb, Ill. : Northern Illinois University Press, 1987. [Carlsson 1992] G. Carlsson, J. Sundberg; "Formant frequency tuning in singing", Journal of Voice, vol. 6, no.3, pp. 256-260, 1992. [Bennet 1989] G. Bennett, X. Rodet; "Synthesis of the Singing Voice", in Current Directions in Computer Music Research, ed. M.V. Mathews & J.R. Pierce, MIT Press, 1989. [Macon 1997a] M.W. Macon et al; "Concatenation-based MIDI-to-Singing Voice Synthesis", Proc. of the 103rd Meeting of Audio Engineering Society, AES Preprint 4591, 1997. [Macon 1997b] M. W. Macon, L. Jensen-Link, J. Oliverio, M. Clements, and E. B. George; "A system for singing voice synthesis based on sinusoidal modeling", Proc. of International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 435-438, 1997. [Peeters 2001] G. Peeters; “Modèles et modélisation du signal sonore adaptes a ses caractéristiques locales” ,Thèse, Ircam, Paris, France, Universite Paris VI 2001. [Rodet 2002] X. Rodet; « Synthesis and Processing of the Singing Voice », 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002), Leuven, 2002. [Bondada 2003] J. Bonada, A. Loscos; "Sample-based Singing- voice Synthesizer by Spectral Concatenation", Proc. of SMAC 03, 439-442, 2003. [Saitou 2004] T. Saitou, N. Tsuji, M. Unoki, M. Akagi; «Analysis of Acoustic Features Affecting" Singing-ness" and Its Application to Singing-Voice Synthesis from Speaking-Voice», Proc. ICLSP, Vol. III, pp. 1929-1932, 2004. [Saitou 2005] T. Saitou, M. Unoki, M. Akagi; « Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis », Speech Communication, Vol 46 : 3-4,pp. 405-417, 2005. [Saitou 2007] T. Saitou, M. Goto, M. Unoki, M. Akagi; « Speech-to-Singing synthesis : converting speaking voices to singing voices by controlling acoustical features unique to singing voices », IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2007. [Banno 2007] H. Banno et. al.; « Implementatioin of realtime STRAIGHT speech manipulation system: Report on its first implémentation », Journal on Acoustic Science and Technology, Vol 28, no 3, pp. 140-146, 2007. [Roebel 2007] A. Roebel, et. al.; « On Cepstral and All-Pole based Spectral Envelope Modeling with unknown Model order”, Pattern Recognition Letters, vol. 28, n° 11, Août, 2007. [Kenmochi 2007] H. Kenmochi, H. Ohshita; "VOCALOID-commercial singing synthesizer based on sample concatenation", Eighth Annual Conference of the International Speech Communication Association, 2007. [Kawahara 2008] H. Kawahara et. al; « TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation », Proc. Conf. on ASSP (ICASSP'08), pp. 3933--3936, 2008. [Veaux 2008] C. Veaux, G. Beller, X. Rodet; "IrcamCorpusTools: an extensible platform for speech corpora exploitation", Proc. of the International Conference on Language Resources and Evaluation (LREC), 2008. [Degottex 2010] G. Degottex; "Glottal source and vocal-tract separation". PhD thesis, UPMC-Ircam, 2010. [Roebel 2010], A. Roebel; « Shape-invariant speech transformation with the phase vocoder », Proc. ICLSP, 2010. [Villavicencio 2010] F. Villavicencio « Conversion de la voix de haute qualité », Université Paris 6 (UPMC), 2010. [Lanchantin 2010] P. Lanchantin, X. Rodet; « Dynamic Model Selection for Spectral Voice Conversion », InterSpeech, Makuhari, 2010. [Roebel 2012] A. Roebel et al, "Analysis and modification of excitation source characteristics for singing voice synthesis", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012.