Vers une mesure automatique de l'adaptation prosodique en interaction conversationnelle

Céline De Looze[1] Stefan Scherer[2] Brian Vaughan[1] Nick Campbell[1]
(1) Speech Communication Lab, Trinity College Dublin, Dublin 2, Irlande
(2) ICT, University of Southern California, Playa Vista, CA, 90094, California
deloozec@tcd.ie, stefan.scherer@gmail.com, bvaughan@tcd.ie, nick@tcd.ie

RÉSUMÉ

Il a été observé dans de nombreuses études qu'un locuteur, au cours d'une conversation, adapte son comportement verbal et non-verbal (lexique, syntaxe, prosodie, postures, gestuelle) à celui de son interlocuteur. Cette adaptation inter-personnelle participe d'une part à faciliter l'échange d'information, la compréhension mutuelle entre interactants et l'atteinte d'un terrain commun. D'autre part, elle augmente chez les acteurs le sentiment d'une interaction sociale réussie en termes de rapport (i.e. relation harmonieuse et attention mutuelle) et d'appartenance sociale. Si l'adaptation inter-personnelle est un phénomène omniprésent de l'interaction conversationnelle, peu de systèmes automatiques et de métriques ont été développés pour la quantifier. Dans cet article, nous présentons un modèle qui permet de mesurer automatiquement l'adaptation prosodique et ses dynamiques en conversation. Sur la base de ce modèle, nous discutons les différentes formes et les dynamiques de l'adaptation prosodique mesurées à partir de conversations téléphoniques enregistrées sur une période de plusieurs mois.

ABSTRACT

Automatic measurement of prosodic accommodation in conversational interaction

It has been observed in many studies that speakers, over the course of a conversation, adapt their verbal and non-verbal behaviour (lexicon, syntax, prosody, postures, gesture) to their interlocutor. This accommodation facilitates, on the one hand, the exchange of information, mutual understanding between interactants and the reaching of common ground. Moreover, it increases the social success of the interaction in terms of rapport (i.e. harmonious relation and mutual attention) and affiliation. While accommodation is a ubiquitous component of social interaction, few automatic systems and metrics have been developed to quantify it. In this paper, we present a model which provides metrics for the automatic measurement of prosodic accommodation and its dynamic manifestation in conversation. Based on this model, we discuss the different forms and the dynamics of prosodic accommodation, measured from conversations recorded over a period of several months.

MOTS-CLÉS : Adaptation prosodique, dynamiques de la parole, interaction sociale.

KEYWORDS: Prosodic adaptation, speech dynamics, social interaction.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 321–328,
Grenoble, 4 au 8 juin 2012.

1 Introduction

De nombreux systèmes de dialogue ont été développés ces dernières années et sont aujourd'hui largement utilisés dans de nombreux domaines tels que la téléphonie mobile, les jeux vidéos ou encore les technologies d'assistance pour les personnes âgées ou handicapées. Si ces systèmes sont capables de traiter la composante linguistique de la communication humaine, ils ne peuvent en revanche toujours pas traiter les dynamiques complexes et les ajustements inter-locuteurs qu'implique l'interaction. Il a été observé dans de nombreuses études qu'un locuteur, au cours d'une conversation, adapte son comportement verbal et non-verbal (lexique, syntaxe, prosodie, postures, gestuelle) à celui de son interlocuteur (Giles *et al.*, 1991; Brennan, 1996; Coulston *et al.*, 2002; Richardson *et al.*, 2007). Cette adaptation inter-personnelle participe d'une part à faciliter l'échange d'information, la compréhension mutuelle entre interactants et l'atteinte d'un terrain commun (Pickering et Garrod, 2004). D'autre part, elle augmente chez les acteurs le sentiment d'une interaction sociale réussie en termes de rapport (i.e. relation harmonieuse et attention mutuelle) et d'appartenance sociale (Tickle-Degnen et Rosenthal, 1990; Duncan *et al.*, 2007). Parce qu'elle joue un rôle important dans l'élaboration du sens mais aussi dans l'expression et la reconnaissance des intentions et états sociaux, son implémentation dans des systèmes existants améliorerait leur efficacité et pourrait faire d'un robot ou d'un avatar un interactant socialement compétent.

Si l'adaptation inter-personnelle est un phénomène omniprésent de l'interaction conversationnelle et a été largement étudiée [1], peu de systèmes automatiques et de métriques ont cependant été développés pour la quantifier. Dans cet article, nous présentons un modèle qui permet de mesurer automatiquement l'adaptation prosodique et ses dynamiques en conversation. Sur la base de ce modèle, nous discutons les différentes formes et les dynamiques de l'adaptation prosodique mesurée à partir de conversations téléphoniques enregistrées sur une période de plusieurs mois.

2 Mesure automatique de l'adaptation prosodique

2.1 Définition d'états

Nous avons proposé dans De Looze et Rauzy (2011) que l'adaptation prosodique (figure 1) peut être décrite au travers d'un ensemble d'états, regroupés autour de trois catégories : l'adaptation, la différentiation et le maintien (cf. la *Communication Accommodation Theory* (Giles *et al.*, 1991)). Dans notre définition, ces catégories sont subdivisées en deux états distincts : la convergence et la synchronie. Lorsque les interactants adoptent un comportement commun, formé au travers des caractéristiques intrinsèques et personnelles de chacun, l'adaptation est convergente. Lorsque les locuteurs coordonnent temporellement les changements ou variations de leur comportement et que ces variations évoluent dans la même direction, l'adaptation est synchrone. En termes de prosodie, une adaptation convergente est par exemple observée lorsque deux locuteurs adoptent un débit de parole similaire ; une adaptation synchrone lorsque deux locuteurs accélèrent et ralentissent leur débit de parole au "même moment" (sujet à décalage temporel du fait de l'organisation des tours de parole). Dans la même veine, la divergence et la synchronie symétrique sont les états de la différentiation. Une divergence peut-être par exemple observée lorsque deux locuteurs exagèrent leurs caractéristiques prosodiques intrinsèques de manière à accentuer leurs différences ; une synchronie symétrique lorsque les variations prosodiques évoluent vers des

1. cf. dans la littérature anglophone les termes *alignement* (Pickering et Garrod, 2004), *convergence* (Giles *et al.*, 1991), *entrainment* (Brennan, 1996), *cameleon effect* (Chartrand et Bargh, 1999), ou encore *mimicry* (Meltzoff et Moore, 1977).

directions opposées (i.e. vers un débit plus rapide vs vers un débit plus lent). Nous émettons l'hypothèse que ces états peuvent être observés individuellement ou en combinaison, ce qui donne un ensemble de 7 états possibles.

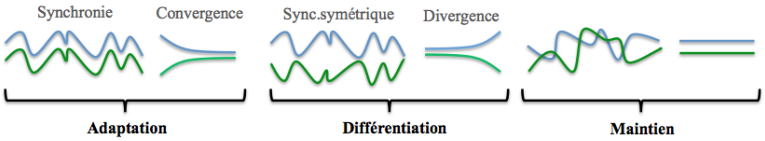

FIGURE 1 – Adaptation prosodique : états

2.2 Extraction des caractéristiques prosodiques

Mesurer automatiquement l'adaptation prosodique entre locuteurs nécessite de définir en premier lieu un domaine ou empan temporel à partir duquel les caractéristiques prosodiques de chaque locuteur seront extraites. Le choix doit se porter vers un empan qui permet une comparaison pertinente des caractéristiques prosodiques des interlocuteurs. La difficulté qui se pose est que leur parole n'est pas alignée temporellement, les locuteurs s'exprimant tour à tour.

Deux méthodes ont été proposées pour l'extraction des caractéristiques prosodiques : la méthode basée sur les tours de parole (*turn-based* ou *utterance-based* ; ex : Levitan et Hirschberg (2011)) et la méthode TAMA (*Time Aligned Moving Average* ; Kousidis *et al.* (2008)). La méthode basée sur les tours de parole consiste à comparer les caractéristiques prosodiques des interlocuteurs tour à tour. L'unité de construction de tour du locuteur A est ainsi comparée à l'unité de construction de tour suivante du locuteur B. Cette méthode présuppose que l'adaptation prosodique se fait très localement, où la production du locuteur A influence directement et uniquement la production consécutive du locuteur B. On peut cependant supposer que l'adaptation prosodique, du fait des dynamiques complexes qu'implique l'interaction, s'effectue sur un empan temporel plus large. Extraire les caractéristiques prosodiques sur chaque tour de parole et mener une comparaison à partir de tours consécutifs uniquement ne parait donc pas une unité pertinente pour mesurer l'adaptation prosodique entre deux locuteurs. Une solution possible est d'étendre cet empan temporel à plusieurs tours de parole comme cela a été suggéré par Nishimura *et al.* (2008). Une autre solution est de choisir une fenêtre temporelle fixe qui recouvre les paroles des deux locuteurs, comme dans la méthode TAMA. La méthode TAMA ne présuppose pas d'empan temporel pour lequel l'adaptation inter-personnelle s'établit. Les caractéristiques prosodiques sont extraites à partir de fenêtres fixes glissantes de durée constante qui se chevauchent en fonction d'un pas d'analyse pré-déterminé. Une telle méthode permet d'obtenir une mesure des indices prosodiques pour chaque locuteur à des intervalles réguliers qui correspondent à un même empan temporel pour les deux locuteurs. Si cette méthode est efficace car elle ne présuppose pas de domaine temporel pour l'adaptation prosodique, elle coupe en revanche de façon aléatoire les productions orales des locuteurs.

Dans notre modèle, nous proposons une méthode hybride inspirée de ces deux méthodes. Nous utilisons comme pour la méthode TAMA un ensemble de fenêtres glissantes qui se chevauchent pour l'extraction des indices prosodiques. A l'instar de la méthode TAMA, les fenêtres glissantes par défaut fixes sont étendues aux bornes de la première et de la dernière unité de construction de tour qu'elles chevauchent. La figure 2 fournit une représentation graphique de ces trois

méthodes. Dans cette étude, la durée de la fenêtre a été fixée à 20 secondes et le pas d'analyse à 10 secondes ; une valeur prosodique pour chaque locuteur est donc extraite toutes les 10 secondes. Les valeurs obtenues sont fonction de la durée de l'énoncé considéré, elles correspondent donc à des moyennes pondérées.

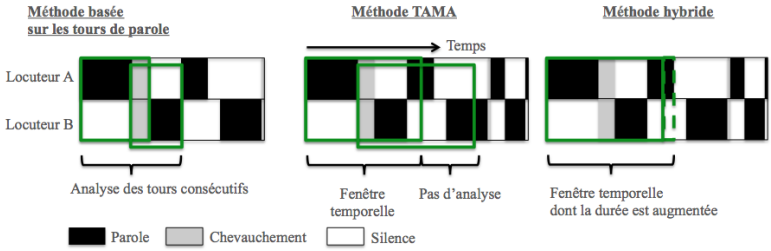

FIGURE 2 – Représentation graphique de la méthode basée sur les tours de parole, de la méthode TAMA et de la méthode hybride.

2.3 Mesures prosodiques

Le modèle extrait un ensemble de paramètres acoustiques à partir des logiciels Praat et MatLab. Ces paramètres rendent compte du registre, de l'intensité de voix et du débit d'élocution des locuteurs.

– registre : médiane (med-f0) et écart type (sd-f0) de la fréquence fondamentale
– intensité : médiane (med-Int) et écart type (sd-Int) de la courbe d'intensité
– débit d'élocution : nombre de syllabes par seconde (syllsec)

Nous avons utilisé une méthode basée sur les modulations à long terme de l'énergie et des caractéristiques spectrales (Maganti *et al.*, 2007) pour une segmentation automatique en intervalles sonores et silencieux. Les noyaux syllabiques ont été automatiquement annotés à partir de l'algorithme de De Jong et Wempe (2009).

2.4 Quantification de l'adaptation prosodique

Dans ce modèle, la *synchronie* est mesurée à partir du coefficient de corrélation linéaire de Bravais-Pearson $\rho_{xy} \in [-1,1]$ qui mesure les dépendances linéaires entre deux ensembles d'observations x and y :

$$\rho_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)\sum_{i=1}^{N}(y_i - \mu_y)}{(N-1)s_x s_y}, \quad (1)$$

où $|x| = |y| = N$, μ_x la valeur moyenne de x (respectivement μ_y) , s_x l'écart type de x (respectivement s_y), and $x_i \in x \quad \forall i = 1,...,N$ (respectivement y_i). Lorsque $\rho_{xy} >> 0$ et proche de 1, la synchronie est très forte ; lorsque $\rho_{xy} << 0$ et proche de -1 la synchronie symétrique est très forte ; lorsque ρ_{xy} est proche de zéro, on n'observe aucune forme de synchronie.

La *convergence* est mesurée à partir de l'intersection de droites linéaires ajustées aux paramètres prosodiques extraits pour chaque locuteur. Pour chaque ensemble de paramètres, l'équation suivante est donnée, où i est l'identifiant du locuteur :

$$y = \alpha_i x + \beta_i, \quad \forall i \in \{1, 2\}, \tag{2}$$

Pour trouver le point d'intersection, on suppose l'égalité des deux équations pour $i \in 1, 2$ (équation 3), ce qui mène après conversion à l'équation 4 :

$$\alpha_1 x + \beta_1 = \alpha_2 x + \beta_2, \tag{3}$$

$$x = \frac{\beta_2 - \beta_1}{\alpha_1 - \alpha_2} \tag{4}$$

Si x, à savoir le point d'intersection, est positif, les deux locuteurs convergent ; si x est négatif, leurs caratéristiques prosodiques divergent. De plus, la valeur x indique la vitesse de convergence (ou de divergence) à partir de laquelle nous estimons la *direction* ou la *force de convergence* de chaque locuteur : si x est proche de zéro la vitesse de convergence est rapide : il y a donc une forte convergence de la part de l'interlocuteur. Si x est plutôt loin de zéro, la vitesse est très lente : le locuteur ne converge que très peu vers son interlocuteur.

2.5 Empan temporel de mesure

Dans de nombreuses études, le phénomène d'adaptation a été investigué en supposant qu'il augmente linéairement au cours du temps. Or, ce qui fait d'une conversation un dialogue interactif, ce sont les changements dynamiques impliqués dans l'interaction. On peut donc supposer que l'adaptation prosodique varie au cours du temps, fonction par exemple de l'engagement des interlocuteurs, comme cela a été observé pour l'anglais dans De Looze et Rauzy (2011) et Vaughan (2011). Afin de mesurer les dynamiques de l'adaptation inter-personnelle, les valeurs de convergence et de synchronie sont extraites dans notre modèle à partir de fenêtres glissantes, similaires à celles utilisées pour la méthode TAMA. Dans notre étude, pour chaque conversation, chaque fenêtre d'analyse correspond à 10 fenêtres d'extraction des caractéristiques prosodiques TAMA [2] ; le pas d'analyse est fixé à 5. La force d'adaptation est donc calculée sur une période de 100 sec. toutes les 50 sec. Pour mesurer l'évolution de l'adaptation au cours de plusieurs conversations, le modèle calule les ratios (ou pourcentages) des états de synchronie et de convergence pour chaque conversation.

3 Données et hypothèses

Pour cette étude, nous avons sélectionné les conversations téléphoniques de 6 locuteurs japonais (3 hommes et 3 femmes formant 4 paires) du corpus JST ESP (Campbell, 2004) ; un total de 40 conversations (10 pour chaque paire et chacune d'une durée de 30 minutes) enregistrées sur une période de plusieurs mois. Pour chaque conversation, les locuteurs étaient libres de parler de ce qu'ils voulaient. Par ailleurs, ils ne se connaissaient pas au début des enregistrements. Ce corpus nous permet de tester deux hypothèses : (1) l'adaptation inter-personnelle est un phénomène dynamique, qui augmente et diminue plusieurs fois au cours du temps ; (2) les 7 états théoriquement définis en 2.1 sont observables en interaction conversationnelle.

2. fenêtres d'extraction décrites en 2.2.

4 Résultats

4.1 Dynamiques intra-conversation

Nos analyses révèlent que pour toutes les conversations, plusieurs phases de synchronie, de synchronie symétrique, de convergence, de divergence et de maintien sont détectées (cf. figure 3). Par ailleurs, les analyses ANOVA (méthode Tukey-Kramer) montrent que le nombre de phases de synchronie pour les paramètres med-f0 et sd-f0 est plus élevé que pour le paramètre syllsec ($p<0.01$). Excepté pour une paire de locuteurs, le nombre de phases de synchronie pour les paramètres med-f0 et sd-f0 est aussi plus élevé que pour le paramètre sd-Int ($p<0.01$). De plus, le nombre de phases de convergence pour les paramètres med-f0 et sd-f0 est plus petit que pour les paramètres syllsec, sd-Int et med-Int ($p<0.01$).

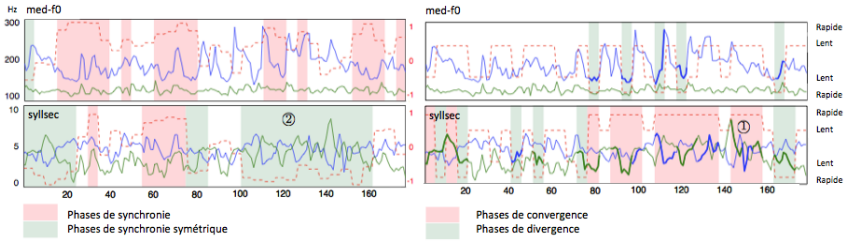

FIGURE 3 –
Données extraites pour chaque locuteur (locuteur 1 en bleu, locuteur 2 en vert) pour chaque fenêtre d'analyse (axe des abscisses) : les deux graphiques du haut représentent les valeurs obtenues pour le paramètre med-f0 (donné en Hz sur l'axe des ordonnées à gauche), les deux graphiques du bas les valeurs obtenues pour le paramètre syllsec (donné en nombre de syllabes par seconde sur l'axe des ordonnées à gauche). Les valeurs de synchronie et de synchronie symétrique sont représentées dans les graphiques de gauche par la ligne en pointillés rouge (valeurs comprises entre -1 et 1, axe des ordonnées à droite). Les phases de synchronie sont colorées en rose, les phases de synchronie symétrique en vert. Les valeurs de convergence et de divergence sont représentées dans les graphiques de droite par la ligne en pointillés rouge (au centre, la convergence/divergence est lente ; aux extrêmes de l'axe des ordonnées, elle est rapide). Les phases de convergence sont colorées en rose, les phases de divergence en vert. Les lignes en gras représentent le locuteur le plus convergeant/divergeant.

4.2 Dynamiques inter-conversations

L'étude de l'évolution des ratios de synchronie/asynchronie et convergence/divergence révèle que le degré d'adaptation inter-personnelle est spécifique à l'interaction. Pour les 4 paires de locuteurs, et pour tous les paramètres prosodiques, ces états varient d'une conversation à l'autre, et ce, de façon non-linéaire.

4.3 Co-occurence des états d'adaptation

L'étude de co-occurence temporelle des états d'adaptation (à partir des graphiques de couleurs en 4) montre que les sept états définis en 2.1 sont observés en interaction conversationnelle. Les états sont observés le plus fréquemment individuellement : l'état de convergence est très peu observé en simultané avec l'état de synchronie ; de même, l'état de divergence est très peu

observé en combinaison avec l'état de synchronie symétrique. Aussi, nous observons que les états de convergence et de synchronie symétrique apparaissent simultanément assez fréquemment pour les paramètres sd-Int et syllsec.

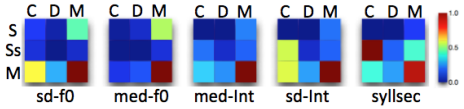

FIGURE 4 – Co-occurences des états de synchronie (S), de synchronie symétrique (Ss), de convergence (C), de divergence (D) et de maintien (M), pour toutes les paires et toutes les conversations ; de gauche à droite pour les paramètres sd-f0, med-f0, med-Int, sd-Int et syllsec. Une couleur chaude réfère à une combinaison très fréquente, une couleur froide à une combinaison peu fréquente.

5 Discussion et conclusion

Dans cet article, nous avons proposé un modèle pour la mesure automatique des dynamiques de l'adaptation prosodique en interaction conversationnelle. Nous avons proposé que l'adaptation prosodique peut être observée et mesurée à partir d'un ensemble de sept états et qu'elle varie au cours du temps (intra- et inter-conversations).

Notre étude montre tout d'abord que les états définis dans notre modèle sont observables à partir de conversations téléphoniques tenues en japonais et que les états les plus fréquents sont ceux observés individuellement. Elle corrobore aussi les observations d'études récentes menées sur l'anglais et confirment que l'adaptation prosodique est un phénomène dynamique : l'adaptation prosodique n'augmente pas linéairement mais varie plusieurs fois au cours d'une conversation. Elle varie aussi au cours du temps (inter-conversations) ce qui suggère qu'elle est plutôt spécifique à l'interaction. Ce travail doit être maintenant complété par plus d'investigations afin de déterminer quelles phases d'adaptation prosodique détectées sont fonctionnellement pertinentes en interaction conversationnelle et quelle(s) méthode(s) (i.e. tours, TAMA, hybride), fenêtre(s) et pas d'analyse (i.e. durée) permettent une description plus fine des dynamiques de l'adaptation prosodique. Notre étude (intra-conversations, section 4.1) révèle par ailleurs que les locuteurs ont tendance à synchroniser les variations temporelles de leur registre plutôt que converger vers des registres similaires ; ils ont au contraire tendance à converger vers un débit de parole similaire plutôt qu'à adapter temporellement leurs variations de débits. Si ces résultats nécessitent plus ample investigation, ils suggèrent que l'adaptation prosodique est contrainte par différents facteurs qui présagent des états à partir desquels elle est observée. On peut par exemple supposer que des contraintes physiques soient la cause d'une adaptation synchrone plutôt que convergente des registres des locuteurs. On peut aussi supposer qu'une perception plus difficile des changements de vitesse d'articulation que des changements de registre se traduise par une vitesse de parole convergente plutôt que synchrone. Cette étude mériterait aussi une investigation systématique du degré d'adaptation de chaque locuteur pour chaque paire, la littérature soulignant une adaptation plus importante chez les femmes et les dyades du même sexe.

Remerciements

Ce travail a été effectué dans le cadre du projet FASTNET - Focus on Action in Social Talk : Network Enabling Technology financé par Science Foundation Ireland (SFI) 09/IN.1/I2631.

Références

BRENNAN, S. (1996). Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, pages 41–44.

CAMPBELL, N. (2004). Speech and expression ; the value of a longitudinal corpus'. *In Proceedings 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 183–186.

CHARTRAND, T. et BARGH, J. (1999). The chameleon effect : The perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893.

COULSTON, R., OVIATT, S. et DARVES, C. (2002). Amplitude convergence in children's conversational speech with animated personas. *In Seventh International Conference on Spoken Language Processing*.

DE JONG, N. et WEMPE, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390.

DE LOOZE, C. et RAUZY, S. (2011). Measuring speakers' similarity in speech by means of prosodic cues : methods and potential. *In Proceedings of Interspeech 2011*, pages 1393–1396. ISCA.

DUNCAN, S., FRANKLIN, A., PARRILL, F. et WELJI, H. (2007). Cognitive processing effects of social resonance in interaction. *Proceedings Gesture 2007-The Conference of the International Society of Gesture Studies, Evanston, IL*.

GILES, H., COUPLAND, N. et COUPLAND, J. (1991). Accommodation theory : Communication, context, and consequence. *Contexts of accommodation : Developments in applied sociolinguistics*, pages 1–68.

KOUSIDIS, S., DORRAN, D., MCDONNELL, C. et COYLE, E. (2008). Times series analysis of acoustic feature convergence in human dialogues. Interspeech.

LEVITAN, R. et HIRSCHBERG, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions.

MAGANTI, H., MOTLICEK, P. et GATICA-PEREZ, D. (2007). Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms. *In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1037. IEEE.

MELTZOFF, A. et MOORE, M. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198(4312):75.

NISHIMURA, R., KITAOKA, N. et NAKAGAWA, S. (2008). Analysis of relationship between impression of human-to-human conversations and prosodic change and its modeling. *In Ninth Annual Conference of the International Speech Communication Association*.

PICKERING, M. et GARROD, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–190.

RICHARDSON, M., MARSH, K., ISENHOWER, R., GOODMAN, J. et SCHMIDT, R. (2007). Rocking together : Dynamics of intentional and unintentional interpersonal coordination. *Human Movement Science*, 26(6):867–891.

TICKLE-DEGNEN, L. et ROSENTHAL, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1(4):285–293.

VAUGHAN, B. (2011). Prosodic Synchrony in Co-operative Task-based Dialogues : A Measure of Agreement and Disagreement. *In Proceedings of Interspeech 2011*, pages 1865–1868. ISCA.