Modélisation de la coarticulation multimodale : vers l'animation d'une tête parlante intelligible

Théo BIASUTTO-LERVAT

Mots-clés – Niveaux et Public concernés

Mots-clés : coarticulation, parole audiovisuelle, *deep learning*, apprentissage des langues

Niveaux : tous niveaux

Public : professeur·e·s de langues étrangères

À quelles questions cette étude tente-t-elle de répondre ?

Dans ces travaux, nous cherchons à simuler par un algorithme l'influence d'un phonème sur la production des phonèmes voisins, phénomène connu sous le nom de coarticulation, dans le but de prédire les mouvements articulatoires nécessaires à la production d'une séquence phonétique. Il faut en effet savoir que nous ne pouvons considérer l'articulation comme une simple concaténation de mouvements articulatoires correspondant aux phonèmes. Dans le domaine d'étude de la production de la parole, il est bien établis que la production d'un phonème est largement influencé par son contexte. Par exemple, la forme des lèvres est très différentes pendant la production du /k/ de « qui » et « quoi », car ce phonème subis une influence de la voyelle suivante. *In fine*, notre modèle permet de contrôler une tête parlante virtuelle, afin de synchroniser son animation à un segment de parole prononcé par un adulte.

Nous cherchons ici à proposer un nouveau modèle de coarticulation basé sur les techniques récentes d'intelligence artificielle (*deep learning*) avec pour principal objectif une modélisation indépendante de la langue et de la modalité. Par modalité, nous entendons l'aspect visuelle et articulatoire de la parole, c'est-à-dire les mouvements du visage induits par l'articulation d'un locuteur (modalité visuelle), mais aussi les mouvements des principaux articulateurs internes comme la langue, la mâchoire ou le vélum (la modalité articulatoire). Finalement, nous souhaitons appliquer ce modèle de coarticulation à la langue Allemande afin de proposer un système de synchronisation labiale automatique pour cette langue, capable d'animer un visage virtuelle depuis la voix de l'enseignant.

Pourquoi ces questions sont-elles pertinentes ?

De nombreuses études ont démontré l'importance de l'information visuelle pour la perception de la parole. En plus de permettre la communication d'informations de haut niveau comme les émotions ou la métacognition (Granström et House, 2005 ; Swerts et Krahmer, 2005), il a été établi que lorsque le signal acoustique est dégradé, la modalité visuelle apportée par le visage peut rétablir jusqu'à deux tiers de l'intelligibilité apportée par l'audio (Le Goff *et al.*, 1994 ; Sumby et Pollack, 1954). Pour de nombreuses applications exploitant actuellement des technologies de synthèse de la parole, l'ajout de la modalité visuelle par le biais d'une tête parlante virtuelle permettrait donc d'augmenter l'intelligibilité de la parole synthétique, et ce, même en l'absence de modélisation interne telle que les mouvements de la langue (Ouni *et al.*, 2007). L'ajout de la modalité visuelle est cependant une tâche critique, car animer un visage virtuel peut se faire au détriment de l'intelligibilité si le signal visuel n'est pas parfaitement congru au signal acoustique. En effet, lorsque nous observons un locuteur, nous utilisons un système neurologique de décodage multimodal, où la modalité visuelle influence notre compréhension de la modalité acoustique, et inversement (Benoit *et al.*, 2010 ; Skipper *et al.*, 2007). Ce mécanisme entraine une plus grande robustesse de la parole aux perturbations extérieures de par la redondance des informations au niveau visuel et acoustique, mais engendre néanmoins une grande sensibilité de l'humain à la moindre incohérence entre les deux modalités de la parole. Que ces incohérences soient dues à une mauvaise synchronisation entre le flux audio et visuel (Dixon et Spitz, 1980), ou à une distorsion phonétique (Green et Kuhl, 1989 ; Green et Kuhl, 1991 ; Jiang *et al.*, 2002), celles-ci peuvent aboutir à d'importants effets sur la perception. L'exemple le plus notable est certainement l'effet McGurk (McGurk et MacDonald, 1976) : quand le stimulus audio 'ba' est couplé à un stimulus visuel 'ga', l'auditeur rapporte entendre prononcer 'da'.

Malgré ces difficultés, le développement de technologies de synthèse audiovisuelle de la parole pourrait être crucial pour la communauté malentendante qui exploitent bien plus le signal visuel qu'une vaste majorité de la population (Campbell *et al.*, 1998 ; MacSweeney *et al.*, 2002), mais il pourrait également être d'une utilité plus générale, adapté à des lieux bruyants, comme les gares ou aéroports. De plus, l'intégration d'avatar virtuel doué de parole peut améliorer l'expérience de l'utilisateur dans de nombreux cadres, comme les assistants virtuels, les sites internet ou les médias sociaux (Cosatto *et al.*, 2003 ; Gibbs *et al.*, 1993). Dans le secteur du divertissement, la synthèse audiovisuelle de la parole pourrait considérablement accélérer la réalisation de film d'animation en automatisant la production d'animation liée à la parole, et il en va de même pour l'industrie vidéoludique.

Enfin, la synthèse audiovisuelle de la parole peut également être utilisée à des fins pédagogiques pour aider à capter l'attention de l'apprenant (Johnson *et al.*, 2000), ou à l'apprentissage de la prononciation des langues étrangères (Hazan *et al.*, 2005 ; Massaro, 2003). Cette utilisation de la synthèse audiovisuelle pour l'apprentissage des langues étrangères peut également être étendue au domaine médical, principalement comme outil de démonstration et de visualisation en orthophonie. En plus d'améliorer la capacité à transmettre des informations, qui est sans conteste l'objectif premier de la parole, l'utilisation d'un visage virtuel capable de parler rend l'interaction avec la machine plus naturelle (Pandzic *et al.*, 1999 ; Sproull *et al.*, 1996), ce qui renforce le confort de l'utilisateur (Dehn et Van Mulken, 2000) et sa confiance dans le système (Ostermann et Millen, 2000). Les utilisateurs interagissant avec un système informatique par le biais d'une tête parlante réagissent donc plus positivement (Pandzic *et al.*, 1999) et sont plus engagés (Sproull *et al.*, 1996 ; Walker *et al.*, 1994).

Quelle méthodologie de recherche a-t-on utilisée ?

Deux grandes étapes ont été au coeur de notre modélisation de la coarticulation.

Dans un premier temps, nous avons procédé à la phase de récolte de données, à l'aide d'outils de capture de mouvement (caméra OptiTrack et articulographie électromagnétique). Ces outils permettent

d'obtenir de très fine mesure de la dynamique articulatoire d'un locuteur, mais on cependant quelques contraintes techniques nous empêchant l'acquisition de nombreuses heures de parole. Afin de s'assurer une bonne qualité des données ainsi qu'une bonne couverture des différents effets de coarticulation, nous avons donc préparé en amont un corpus textuel nous assurant une excellente richesse phonétique tout en minimisant la quantité de phrases à enregistrer.

Dans un second temps, nous avons procédé à la phase de conception et de validation de notre modèle de coarticulation. Cette phase est elle-même divisée en deux sous-étapes : une étape dite *exploratoire* consistant en l'élaboration d'un réseau de neurones artificiels capable d'apprendre la dynamique des articulateurs en fonction d'une séquence phonétique, et une étape dite *d'évaluation subjective* servant à valider notre modèle une fois appliqués à un système d'animation facial.

L'étape *exploratoire* fut guidée par l'utilisation de métrique usuelle du domaine (calculs de l'Erreur Quadratique Moyenne et de corrélations) servant à mesurer l'erreur commise par notre modèle sur sa prédiction de la dynamique des articulateurs par rapport aux données précédemment acquises (c'est-à-dire, au cours de la phase de récolte de données), ainsi que par une analyse fine de cette dynamique, afin de mesurer la qualité de nos prédictions vis-à-vis des cibles articulatoires critiques pour l'intelligibilité de la parole. Par exemple, cette analyse a été appliquée aux consommes bilabiales /b/, /p/ et /m/, pour lesquelles il est indispensable d'obtenir une fermeture totale des lèvres durant la production, faute de quoi le phonème perçu sera différent (voir effet McGurck à la section précédente).

L'étape *d'évaluation subjective* a été réalisée via une plateforme web, afin de mesurer la différence de préférence entre un visage 3D animé depuis les données de capture de mouvement (issues de notre acquisition de données), et un visage 3D animé par notre modèle de coarticulation. Dans ces deux cas de figures, la voix utilisé est celle enregistré pendant la capture de données. Afin d'évaluer la pertinence de notre modèle, nous avons proposés 50 paires de vidéos à 10 natifs Allemands adultes, ces vidéos couvrant un très grand ensemble des phonèmes de la langue Allemande. Pour chaque paires de vidéos, les participants ont le choix entre 6 niveaux afin de préciser leurs préférences pour l'un ou l'autre des échantillons (préférence pour : "A", "plutôt A", "un peu A", "un peu B", "plutôt B", "B"). Bien entendu, l'ordre d'apparition des vidéos est entièrement aléatoires.

Quels résultats a-t-on obtenus ?

Dans cette étude, nous avons validé l'utilisation d'un modèle particulier de réseaux de neurones pour la modélisation de la coarticulation, les réseaux de neurones récurrents bidirectionels. Ces derniers sont capables de prendre en compte l'information passée et future pour générer de la dynamique des articulateurs, capacité permettant de prendre en compte les phénomènes de coarticulation rétentive (l'influence des phonèmes passés sur la production du phonème courant) et de coarticulation anticipatoire (l'influence des phonèmes futures sur la production du phonème courant).

En termes de mesures objectives (erreur moyenne et corrélation), nos résultats atteignent l'état de l'art du domaine, avec cependant l'avantage majeur de pouvoir être appliqué à n'importe quelle langue et articulateur. Par exemple, nous avons utilisé ce modèle dans cette étude pour prédire la modalité visuelle (les déformations du visage) mais aussi la modalité articulatoire (la dynamique des principaux articulateurs internes comme la langue ou le palais mou), et ce dans différentes langues (Anglais, Allemand et Français). En comparant nos résultats sur un corpus articulatoire ouvert avec ceux issues de l'inversion acoustique, tâche consistant en la prédiction des trajectoires articulatoires depuis le signal acoustique, nous avons également validé que l'utilisation d'informations phonétiques seules (phonème et durée respective) est suffisante pour prédire la modalité articulatoire avec une grande précision.

Concernant l'évaluation subjective, il peut sembler très difficile de comparer nos prédictions aux données issues de la capture de mouvement. En effet, il semble impossible de prédire des trajectoires articulatoires « meilleures » que les données originales. L'objectif de notre évaluation est donc de montrer qu'il est très difficile pour les participants de différencier les deux animations. Nous avons pour cela développé un simple taux d'appréciation basé sur le choix catégorique de préférence des

participants. Une valeur de 50 % de ce taux représente donc une appréciation égale entre nos prédictions et la capture de mouvement. Nous avons obtenu un honorable score de 41 %, signifiant donc une légère préférence pour les trajectoires articulatoires originales. Une analyse fine de ces résultats révèle que certaines prédictions sont très nettement rejetées par l'utilisateur (c'est-à-dire, nos 10 participants), ouvrant ainsi des pistes de réflexion intéressantes pour l'amélioration de ce modèle.

Que dois-je retenir de cette étude pour ma pratique ?

- La parole ne se limite pas à une onde sonore, et les informations visuelles apportées par le visage sont utilisées par l'interlocuteur pour décoder le message.
- L'information visuelle augmente l'intelligibilité du discours, particulièrement lorsque le signal acoustique est dégradé.
- Cette intelligibilité supplémentaire peut-être restaurée à l'aide d'outils numériques.
- L'utilisation d'une tête parlante virtuelle aide à capter l'attention et augmenter la compréhension des apprenants.

Références

Benoit, M. M., Raij, T., Lin, F.-H., Jääskeläinen, I. P. et Stufflebeam, S. (2010). Primary and multisensory cortical activity is correlated with audiovisual percepts. *Human Brain Mapping*, *31*(4), 526-538.

Campbell, R., Burnham, D., Dodd, B., Campbell, R., Away, G. et Burnham, D. K. (1998). *Hearing by eye II : Advances in the psychology of speechreading and auditory-visual speech* (vol. 2). Psychology Press.

Cosatto, E., Ostermann, J., Graf, H. P. et Schroeter, J. (2003). Lifelike talking faces for interactive services. *Proceedings of the IEEE*, *91*(9), 1406-1429.

Dehn, D. M. et Van Mulken, S. (2000). The impact of animated interface agents : a review of empirical research. *International Journal of Human-Computer Studies*, *52*(1), 1-22.

Dixon, N. F. et Spitz, L. (1980). The detection of audiovisual desynchrony. *Perception*, *9*, 719-721.

Gibbs, S., Breiteneder, C., De Mey, V. et Papathomas, M. (1993). Video widgets and video actors. Dans *Proceedings of the 6th annual ACM symposium on User interface software and technology* (p. 179-185).

Granström, B. et House, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, *46*(3-4), 473-484.

Green, K. P. et Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception and Psychophysics*, *45*, 34-42.

Green, K. P. et Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology : Human Perception and Performance*, *17*, 278-288.

Hazan, V., Sennema, A., Iba, M. et Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, *47*(3), 360-378.

Jiang, J., Alwan, A., Keating, P. A., Auer, E. T. et Bernstein, L. E. (2002). On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. *EURASIP Journal on Applied Signal Processing*, *11*, 1174-1188.

Johnson, W. L., Rickel, J. W., Lester, J. C. *et al.*, (2000). Animated pedagogical agents : Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, *11*(1), 47-78.

Le Goff, Guiard-marigny, T., Cohen, M. et Benoit, C. (1994). Real-Time Analysis-Synthesis and Intelligibility of Talking Faces. Dans *2nd International conference on Speech Synthesis* (p. 53-56).

MacSweeney, M., Calvert, G. A., Campbell, R., McGuire, P. K., David, A. S., Williams, S. C., Woll, B. et Brammer, M. J. (2002). Speechreading circuits in people born deaf. *Neuropsychologia*, *40*(7), 801-807.

Massaro, D. W. (2003). A computer-animated tutor for spoken and written language learning. Dans *Proceedings of the 5th international conference on Multimodal interfaces* (p. 172-175).

McGurk, H. et MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.

Ostermann, J. et Millen, D. (2000). Talking heads and synthetic speech : An architecture for supporting electronic commerce. Dans *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)* (vol. 1, p. 71-74). IEEE.

Ouni, S., Cohen, M. M., Ishak, H. et Massaro, D. W. (2007). Visual contribution to speech perception : measuring the intelligibility of animated talking heads. *EURASIP Journal on Audio, Speech, and Music Processing*, *2007*(1), 3-3. https://doi.org/10.1155/2007/47891

Pandzic, I. S., Ostermann, J. et Millen, D. (1999). User evaluation : Synthetic talking faces for interactive services. *The Visual Computer*, *15*(7), 330-340. https://doi.org/10.1007/s003710050182

Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C. et Small, S. L. (2007). Hearing lips and seeing voices : how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, *17*(10), 2387-2399.

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H. et Waters, K. (1996). When the interface is a face. *Human-Computer Interaction*, *11*(2), 97-124.

Sumby, W. et Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212.

Swerts, M. et Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, *53*(1), 81-94.

Walker, J. H., Sproull, L. et Subramani, R. (1994).