Etude comparative des différentes unités acoustiques pour la synchronisation labiale

Yannick BENEZETH, Guillaume GRAVIER, Frédéric BIMBOT

IRISA/CNRS & INRIA,
Campus de Beaulieu, 35042 Rennes, France
yannick.benezeth@inria.fr,frederic.bimbot@irisa.fr, guillaume.gravier@irisa.fr

Résumé – Nous étudions dans cet article les performances d'un système d'animation labiale basé sur différentes unités acoustiques. Nous analysons plus particulièrement l'utilisation du phonème et du visème, avec ou sans leur contexte acoustique, modélisés par des modèles de Markov cachés. L'originalité des travaux présentés dans cet article réside dans l'utilisation d'ensembles de phonèmes et de visèmes adaptés à la langue française et dans l'étude de l'impact de la taille de la table de visèmes sur le choix de l'unité acoustique. Les résultats expérimentaux présentés montrent que de meilleures performances sont obtenues en utilisant les tri-phonèmes lorsque la table des visèmes est réduite et inversement, les tri-visèmes lorsque la table des visèmes est plus précise.

Abstract – In this paper, we study the performance of a lip animation system based on different acoustic units. We analyze in particular the use of phonemes and visemes, with or without the acoustic context, modeled by Hidden Markov Models. The presented work is based on phonemes and visemes sets adapted to the French language. We also study the impact of the visemes set size on the choice of acoustic unit. The experimental results show that better performances are obtained using the tri-phonemes when the viseme set is reduced and, the tri-visemes when the viseme set is more accurate.

1 Introduction

La perception de la parole est multimodale. La composante visuelle est combinée aux informations audio et a une grande influence sur la perception de la parole (notamment en environnement bruité [2]). Cette composante est particulièrement importante pour les personnes sourdes ou malentendantes. La multimodalité de la parole a, entre autres, été démontré dans l'effet McGurk. McGurk et MacDonald [1] ont réalisé une expérience où était présenté un extrait vidéo d'une personne prononçant la syllabe /ga/ en synchronie avec un extrait sonore de la syllabe /ba/ et ont montré que le système perceptif humain reconnaissait un /da/. Une incohérence entre la composante visuelle et l'information visuelle dégrade donc la compréhension d'un discours. De plus, les modalités audio et visuelle peuvent aussi porter des informations complémentaires. Par exemple, la distinction entre la syllabe /ba/ et /ga/ est facilement réalisée à partir de l'information visuelle alors que la distinction entre /ka/ et /ga/ est plus fiable à partir de la modalité audio [3].

Même si l'intégration d'informations hétérogènes est une tâche difficile, les systèmes de reconnaissance de la parole récents utilisent la multimodalité de la parole en intégrant l'analyse visuelle dans le processus de reconnaissance. La fusion de ces modalités hétérogènes peut être réalisée en amont de la chaîne de traitement en fusionnant les descripteurs audio et visuel (*e.g.* dans [4]). La fusion peut également être tardive en intégrant les résultats de la reconnaissance audio et visuelle (*e.g.* dans [5]).

Les problématiques liées à l'animation de personnages virtuels (*e.g.* dans les jeux-vidéos, le cinéma ou les interfaces hommes/machines) considèrent également la multimodalité de la parole, mais sous un angle différent. L'objectif étant souvent de synthétiser les mouvements de lèvres de personnages virtuels à partir d'un flux de parole (réel ou synthétique). Pour que l'animation d'un personnage virtuel soit réaliste, il est nécessaire que les mouvements de ses lèvres soient cohérents avec son discours. Il est possible de réaliser une telle animation manuellement en ajustant trame après trame les paramètres de contrôle des lèvres de l'avatar. Les résultats d'animation sont alors très réalistes, mais ce procédé demande une charge de travail conséquente et ne permet pas des animations d'avatar en temps-réel.

Il existe donc des méthodes pour générer automatiquement des mouvements de lèvres à partir d'un flux de parole, en l'occurrence un signal audio. Celles-ci peuvent être divisées en deux classes. Il y a tout d'abord les méthodes qui établissent directement la correspondance entre des descripteurs audio, *e.g.* Mel Frequency Cepstral Coefficients (MFCC) ou Linear Predictive Coding (LPC), et des paramètres de contrôle de la forme de la bouche (*e.g.* ouverture, écartement, protrusion). Cette correspondance est apprise par exemple avec des réseaux de neurones, des modèles de mélanges de gaussiennes ou une quantification vectorielle (*e.g.* [6, 7]). Une deuxième classe de méthode repose sur le principe de la reconnaissance de la forme de la bouche à partir du signal audio correspondant. Les mouvements de la bouche sont alors contrôlés avec un ensemble dis-

cret de positions : les visèmes. Un visème est donc une forme de bouche associée à des sons particuliers. Une analyse spectrale et/ou temporelle fournit des vecteurs d'observation qui sont ensuite utilisés pour la reconnaissance de la forme de la bouche. La reconnaissance peut être basée sur les techniques de reconnaissance de la parole, *e.g.* avec les modèles de Markov cachés (HMM) [8].

Pour les méthodes de cette deuxième catégorie, l'unité acoustique de la parole modélisé peut être le phonème [9] ou le visème [10]. Si le système de reconnaissance est basé sur la reconnaissance de phonèmes, il faut ensuite faire la correspondance entre chaque phonème détecté et son visème correspondant.

Les travaux présentés dans cet article se situent dans le cadre du projet collaboratif Rev-TV dont l'objectif est de créer une nouvelle catégorie d'émission TV où les téléspectateurs ont la possibilité d'interagir avec le contenu de l'émission au travers d'un environnement immersif et convivial. Les téléspectateurs qui interagissent avec le contenu de l'émission auront leur représentation virtuelle dans un environnement de réalité augmentée. Cet avatar sera contrôlé par plusieurs modalités (analyse des mouvements, du son etc.). Cet article présente les travaux menant à une animation des lèvres du personnage virtuel à partir du signal audio. Nous étudions ici les performances d'un système d'animation labiale basé sur différentes unités acoustiques. Nous analysons plus particulièrement l'utilisation du phonème et du visème, avec ou sans leur contexte acoustique, modélisés par des HMMs.

Dans la suite de cet article, nous présentons tout d'abord le protocole expérimental de notre étude en détaillant les ensembles de phonèmes et de visèmes utilisés, la base sur laquelle l'apprentissage et les tests ont été réalisés ainsi que la méthode de reconnaissance d'unités acoustiques mise en oeuvre dans cette étude. Nous présentons et analysons ensuite les résultats obtenus.

2 Protocole expérimental

L'objectif de notre étude est donc d'étudier les performances d'un système d'animation labiale à partir d'un flux de parole. Nous utilisons les HMMs pour reconnaitre différentes unités acoustiques. Nous comparons plus particulièrement les performances obtenues lorsque les unités acoustiques modélisées sont les phonèmes ou les visèmes. Dans cette partie, nous présentons tout d'abord les différentes unités acoustiques évaluées. Nous présentons ensuite la méthode de reconnaissance de ces unités acoustiques ainsi que la base sur laquelle les expériences ont été menées.

2.1 Les différentes unités acoustiques

Nous utilisons un ensemble de 33 phonèmes pour décrire l'ensemble de la langue française (*cf.* Tables 1 et 2 où le symbole # représente le silence). Nous utilisons également deux ensembles de visèmes, le premier (appelé par la suite *VIS1*) est composé de 17 visèmes (*cf.* Table 1) et est inspiré des travaux de Benoit *et al.* [12]. Le second ensemble de visèmes (appelé par la suite *VIS2*) est composé de 8 visèmes, celui-ci a été établi par Govokhina [13]. Ces deux ensembles de visèmes ont été définis pour la langue française mais présentent des niveaux de précision différents. Par exemple, le visème n°6 de *VIS2* est composé de l'union des visèmes n° 5 et 6 de *VIS1*. Trois exemples de visèmes sont présentés dans la Figure 1.

TABLE 1 – Ensemble des 17 visèmes de *VIS1*.

Visèmes	Phonèmes
0	#
1	a
2	i
3	y, u, ø, œ, o, ɔ̃, ɥ
4	e, ɛ, ɛ̃
5	ɔ
6	ɑ̃
7	p, b, m
8	t, d, n, ɲ
9	k, g
10	f, v
11	s, z
12	ʃ, ʒ
13	l
14	ʁ
15	w
16	j

TABLE 2 – Ensemble des 8 visèmes de *VIS2*.

Visèmes	Phonèmes
0	#
1	p, b, m
2	f, v
3	ʃ, ʒ
4	t, d, n, ɲ, s, z, k, g, l, ʁ
5	e, ɛ, ɛ̃, a, i, j
6	ɔ, ɑ̃
7	y, u, ø, œ, o, ɔ̃, ɥ, w

Afin de prendre en compte le contexte de chaque unité acoustique, c'est à dire l'unité acoustique précédente et suivante, nous considérerons également par la suite des triplets de phonèmes (ou visèmes). Ces unités acoustiques sont appelés triphonèmes ou tri-visèmes.

FIGURE 1 – Exemple de cinq visèmes de l'ensemble *VIS1*.

2.2 Reconnaissance d'unités acoustiques

Le système de reconnaissance d'unités acoustiques utilisé pour cette étude est basé sur les modèles de Markov cachés. L'objectif est de trouver la séquence de phonèmes ou de visèmes la plus vraisemblable étant donnée une séquence d'observations acoustiques. Pour cela, le flux de parole (*i.e.* le signal audio) est tout d'abord décomposé en séquence de vecteurs d'observation. Nous utilisons 39 *MFCC* (13 coefficients statiques et leurs dérivés du premier et second ordre) calculés sur des trames de 20ms d'une période de 10ms.

Nous utilisons des modèles de Markov à trois états avec une topologie gauche-droite où les probabilités d'observation des HMMs sont modélisées par des mélanges de 32 gaussiennes. L'algorithme de Baum-Welch est utilisé pour réestimer les paramètres de chaque HMM et l'algorithme de Viterbi est utilisé pour déterminer la séquence de phonèmes ou de visèmes la plus probable [8]. Cette approche est aujourd'hui couramment utilisée pour la reconnaissance de mots. Si l'unité acoustique utilisée est le phonème, une simple correspondance est ensuite établie pour obtenir la séquence de visèmes à partir des Tables 1 ou 2.

Cependant, cette représentation ne prend pas en compte le contexte de chaque unité acoustique. Pour cela, on utilise les tri-phonèmes ou les tri-visèmes. Si nous avons un ensemble de N phonèmes, nous obtenons logiquement un ensemble de N^3 tri-phonèmes. Dans ce cas, la quantité de données d'apprentissage nécessaire devient trop importante. Pour réduire le nombre de paramètres à estimer, les états avec le même contexte sont liés, c'est à dire qu'ils partagent le même jeu de paramètres du modèle HMM.

L'apprentissage a été réalisée sur le corpus de la campagne ESTER [11]. Ce corpus est composé d'émissions d'information radio et TV dont la transcription phonétique a été annotée. Nous utilisons plus de 100 heures de ce corpus pour l'apprentissage et environ 30 minutes pour les séquences de test.

3 Analyse des résultats

Nous présentons dans cette partie les résultats expérimentaux de notre étude comparative. Les résultats sont présentés en utilisant la *précision* et le *rappel*. La *précision* représente le pourcentage d'unités acoustiques correctement détectées par rapport au nombre total d'unités acoustiques détectées et le *rappel* correspond au pourcentage d'unités acoustiques correctement détectées par rapport au nombre d'unités acoustiques annotées dans la vérité terrain. Nous présentons également la moyenne harmonique de ces deux valeurs définie par :

$$\text{F-score} = 2 * \frac{\text{précision} * \text{rappel}}{\text{précision} + \text{rappel}}. \quad (1)$$

Le *F-score* combine la *précision* et le *rappel* et permet de comparer les résultats des différentes méthodes à partir d'une seule valeur. La valeur du *F-score* varie entre 0 et 1, où 1 est le meilleur score. Nous présentons dans la Table 3 les résultats obtenus en utilisant l'ensemble de visèmes *VIS1* et dans la Table 4 les résultats obtenus avec l'ensemble de visèmes *VIS2*.

TABLE 3 – Résultats obtenus en utilisant *VIS1*.

Unités acoustiques	Précision	Rappel	F-score
Phonèmes	66,18	59,97	62,92
Tri-phonèmes	69,61	64,61	67,02
Visèmes	67,38	62,85	65,04
Tri-visèmes	**70,78**	**69,35**	**70,06**

TABLE 4 – Résultats obtenus en utilisant *VIS2*.

Unités acoustiques	Précision	Rappel	F-score
Phonèmes	75,48	72,05	73,72
Tri-phonèmes	**80,64**	**77,64**	**79,11**
Visèmes	69,66	67,80	68,71
Tri-visèmes	73,03	75,60	74,29

Plusieurs observations peuvent être formulées à partir des résultats présentés ci-dessus. Tout d'abord et sans surprise, la prise en compte du contexte (avec les tri-phonèmes ou les tri-visèmes) augmente nettement les performances de reconnaissance par rapport aux phonèmes et aux visèmes.

Ensuite, il est intéressant de constater que la taille de la table des visèmes a un impact sur les performances. En effet, moins la table des visèmes contient de classes et plus les différents HMMs doivent généraliser un modèle rassemblant des sons va-

riés. L'exemple du visème n°4 de l'ensemble *VIS2* est particulièrement représentatif puisque les paramètres du HMM doivent ici modéliser un ensemble de sons (et donc de vecteurs d'observation) très différents. Cette observation explique également le résultat suivant : lorsque la table des visèmes est réduite, les meilleures performances sont obtenues en utilisant le phonème comme unité acoustique (et *a fortiori* le tri-phonème) et lorsque la table des visèmes est plus précise, les meilleures performances sont obtenues en utilisant le visème. En effet, on peut observer dans les Tables 3 et 4 que les meilleures performances ont été obtenues avec les tri-visèmes pour l'ensemble *VIS1* et avec les tri-phonèmes pour l'ensemble *VIS2*.

Cette remarque vient compléter les résultats présentés dans la littérature où les meilleures performances sont systématiquement obtenues avec les tri-visèmes (*e.g.* dans [14]). En effet, les tables de visèmes utilisées dans ces études sont constitués des 16 visèmes définis dans la norme MPEG4 (adaptés à la langue anglaise).

4 Conclusions et perspectives

Nous avons présenté dans cet article une étude comparative des différentes unités acoustiques dans un contexte d'animation labiale de personnages virtuels. Nous avons particulièrement étudié l'utilisation du phonème et du visème, avec ou sans leur contexte acoustique, modélisés par des HMMs. Les ensembles de phonèmes et de visèmes utilisés sont adaptés à la langue française et l'impact de la taille de la table de visèmes sur le choix de l'unité acoustique a été analysé. Nous avons tout d'abord confirmé quantitativement que la prise en compte du contexte (avec les tri-phonèmes ou les tri-visèmes) augmente nettement les performances de reconnaissance par rapport aux phonèmes et aux visèmes. Ensuite, les résultats expérimentaux montrent que de meilleures performances sont obtenues en utilisant les tri-phonèmes lorsque la table des visèmes est réduite et inversement, les tri-visèmes lorsque la table des visèmes est plus précise.

Est-ce qu'un système d'extraction de visèmes présentant un *F-score* de 0.70 permet d'obtenir des animations labiales réalistes ? Afin de déterminer si les résultats présentés ici sont suffisants pour animer de manière réaliste un personnage virtuel ou si une version dégradée (et donc plus rapide) présente des résultats d'animation satisfaisants, nous travaillerons sur une évaluation subjective venant compléter cette évaluation au niveau symbolique.

Références

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices", Nature, Vol. 264, no. 5588, pp. 748-756, 1976.

[2] D. Reisberg, J. McLean and A. Golffield, "Easy to hear, but hard to understand : a lipreading advantage with intact auditory stimuli", Hearing by Eye : The Psychology of Lipreading, London : Lawrence Erlbaume, pp. 97-113, 1987.

[3] D. W. Massaro and D. G. Stork, "Speech recognition and sensory integration", American Scientist, Vol. 86(3), pp. 236–244, 1998.

[4] M.J. Tomlinson, M.J. Russel and N.M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition", in Proceedings of the IEEE international conference on Acoustic Speech and Signal Processing, pp. 821–824, 1996.

[5] J.-S. Lee and C.-H. Park, "Robust Audio-Visual Speech Recognition Based on Late Integration", IEEE Transactions on Multimedia, Vol. 10(5), pp. 767–779, 2008.

[6] T. Frank, M. Hoch and G. Trogemann, "Automated Lip-Sync for 3D-Character Animation", 15th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics, pp. 24–29, 1997.

[7] S. Nakamura, "Statistical multimodal integration for audio-visual speech processing", IEEE Transactions on Neural Networks, pp. 854–866, Vol. 13(4), 2002.

[8] L.-R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, Vol. 77(2), pp. 257–286, 1989.

[9] J. Park and H. Ko, "Real-Time Continuous Phoneme Recognition System Using Class-Dependent Tied-Mixture HMM With HBT Structure for Speech-Driven Lip-Sync", IEEE Transactions on Multimedia, Vol. 10(7), pp. 1299–1306, 2008.

[10] S.-W. Foo and L. Dong, "Recognition of Visual Speech Elements Using Hidden Markov Models ", Advances in Multimedia Information Processing, Vol. 2532, pp. 153–173, 2002.

[11] G. Gravier, J. Bonastre, E. Geoffrois, S. Galliano, K. McTait et K. Choukri, "ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français", Journées d'Etude sur la Parole (JEP), 2004.

[12] C. Benoit, T. Lallouache, T. Mohamadi et C. Abry, "A set of French visemes for visual speech synthesis", Les cahiers de l'ICP, Rapport de recherche, Vol. 3, pp. 113-129, 1994.

[13] O. Govokhina, "Modèles de trajectoires pour l'animation de visages parlants", Thèse de l'Institut National Polytechnique de Grenoble, 2008.

[14] E. Bozkurt, Q.-E. Erdem, E. Erzin, T. Erdem, M. Ozkan, "Comparison of Phoneme and Viseme Based Acoustic Units for Speech Driven Realistic lip Animation", in 3DTV international Conference, 2007.