Modèles multi-flux asynchrones pour la reconnaissance audio-visuelle de la parole: des chiffres au grand vocabulaire.

Guillaume Gravier *Gerasimos Potamianos* *Chalapathy Neti*

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
Mél: (ggravier, gpotam, cneti)@us.ibm.com

ABSTRACT

We investigate the use of multi-stream HMMs for audio-visual speech recognition. Multi-stream HMMs allow the modeling of asynchrony between the audio and visual state sequences at a variety of levels (phone, syllable, word, etc) and can be seen as a product HMM. In this paper, we use such models to investigate the impact of allowing a controlled level of asynchrony. Furthermore, we investigate joint training of the product HMM parameters, compared to composing the model from separately trained audio- and visual-only HMMs. Experiments are carried out on a simple digit recognition task as well as on a more complex dictation task. Results show that in both cases, joint training outperforms independent training. We also show that asynchrony helps a lot on the digit recognition task while surprisingly, it does not yield any improvement on the dictation task.

1. INTRODUCTION

En dépit des progrès récents en reconnaissance automatique de la parole, les systèmes actuels restent sensibles au canal, à l'environnement et au style d'élocution. Plusieurs solutions, comme la compensation de canal et/ou l'adaptation au locuteur, ont été proposés pour accroître la robustesse des systèmes. Cependant, ces techniques ne sont pas efficaces pour un signal fortement dégradé. Une autre approche du problème consiste à utiliser d'autres sources d'information que le signal lui-même. En particulier, les informations visuelles, comme le mouvement des lèvres, semblent une bonne piste dans la mesure où ces informations sont fortement corrélées au signal de parole et moins sensibles aux bruits et à l'environnement. Il a également été montré que les auditeurs humains s'appuient sur l'information visuelle pour reconnaître la parole.

Ces dernières années ont vu un essor considérable des systèmes de reconnaissance audiovisuels (AV) qui combinent les informations provenant du signal avec celles provenant d'une image du visage ou de la bouche du locuteur [8, 3, 5, 7]. Mis à part le problème de la représentation de l'information visuelle, le problème principal est l'intégration des deux sources d'informations. Les méthodes permettant une telle intégration sont habituellement divisées en deux grandes familles : fusion de représentations et fusion de décisions. La première famille rassemble les techniques utilisant une projection des représentations audio et visuelle dans un espace de représentation audiovisuel avant d'utiliser des modèles classiques, comme les modèles de Markov cachés (MMC), sur l'espace de représentation conjoint [2, 6]. Les méthodes de la deuxième famille s'appuient sur la combinaison de décisions, éventuellement partielles, prises indépendamment sur les informations audio d'une part et visuelles d'autre part.

Le MMC multi-flux (MMC-MF), qui combine linéairement les log-vraisemblances audio et vidéo, est un exemple classique de fusion de décision [3, 5]. Ce modèle, proposé à l'origine pour la reconnaissance de la parole par sous-bandes, a été utilisé avec succès dans le cadre AV et permet de contrôler le degré d'asynchronie entre les deux flux de données. Dans la plupart des cas, les vraisemblances sont combinées au niveau de l'état dans le MMC, forçant ainsi les deux flux à être synchrones. Cependant, bien que les données audio et visuelles soient corrélées, elles ne sont pas synchrones et l'activité visuelle précède souvent le signal sonore. Pour prendre en compte cette asynchronie, les MMC-MF peuvent être utilisés pour combiner les vraisemblances à un niveau plus élevé que la trame, comme le phone ou encore le mot [8, 4, 3, 5]. Les MMC multi-flux asynchrones se basent sur l'utilisation d'un MMC composite construit comme le produit des MMC audio et visuels.

Cet article décrit l'utilisation du modèle produit pour la reconnaissance audiovisuelle de la parole sur une tâche de reconnaissance de chiffres et sur une tâche plus complexe de reconnaissance grand vocabulaire de parole lue. Deux raisons principales président au choix de cette étude. Premièrement, les précédents travaux sur les MMC produits ne considèrent pas l'estimation conjointe des paramètres du modèle produit [8, 4, 3], ou bien la considère en ne respectant pas le partage des paramètres issus de la composition des deux MMC [5]. Nous étudierons donc l'estimation conjointe des paramètres des MMC audio et visuel dans le cadre du modèle produit en respectant le partage des densités conditionnelles dans chacun des flux. Deuxièmement, à la connaissance des auteurs, les modèles produits ont rarement été étudiés sur des grands vocabulaires.

2. LE MODÈLE MULTI-FLUX

2.1. Aspects pratiques et théoriques

Le principe des MMC multi-flux consiste à modéliser chaque flux de manière indépendante à l'aide de MMC entre deux points de synchronisation. Dans cette étude, nous considérons les frontières de phones comme points de synchronisation. A chaque frontière de phone, les log-vraisemblances des deux flux sont combinées linéairement pour obtenir la probabilité a posteriori du phone.

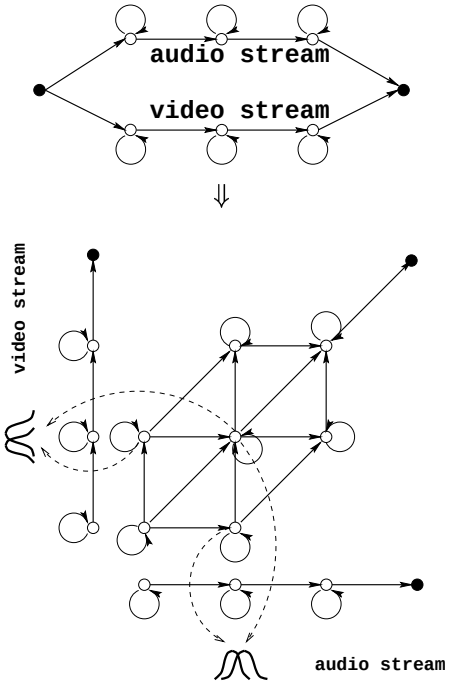


Figure 1: MMC multi-flux et le modèle produit correspondant.

Sans perte de généralité, nous considérerons que les coefficients de la combinaison linéaire ne dépendent pas du phone. Ce modèle peut-être vu comme un produit de MMC comme illustré figure 1. Dans le modèle produit, la log-vraisemblance conditionnelle des données pour l'état i est donnée par

$$\begin{aligned}\ln p(y_a(t), y_v(t)|i) &= \alpha \ln p(y_a(t)|s_a(i)) + \\ &\quad (1-\alpha) \ln p(y_v(t)|s_v(i)) \ ,\end{aligned}$$

où $y_a(t)$ (resp. $y_v(t)$) est le vecteur de données audio (resp. visuelles) à t et $s_a(i)$ (resp. $s_v(i)$) est une fonction qui désigne l'état correspondant à l'état produit i dans le MMC audio (resp. visuel). La figure 1 illustre aussi le partage des densités dans le modèle produit (en bas), les probabilités de transition étant données par le produit des probabilités de transition des MMC unimodaux. Il convient de noter que, bien que les vraisemblances soient combinées au niveau de l'état dans le modèle produit, cela est strictement équivalent à la combinaison aux frontières de phones, ce qui justifie l'appartenance de cette méthode à la famille des méthodes basées sur la fusion de décisions.

De plus, il est aisé de limiter l'asynchronie entre les deux flux de données en restreignant le produit aux seuls états (produits) valides. On mesurera l'asynchronie, notée a, dans un état produit comme la valeur absolue de la différence entre les numéros des états des éléments du produit. La figure 1 (en bas) illustre un modèle admettant une asynchronie $a = 1$, caractérisée par l'absence de deux états par rapport aux 9 états du produit complet. A l'extrême, $a = 0$, le modèle produit est équivalent au modèle multi-flux synchrone classique[1] dans lequel un seul MMC est utilisé pour représenter la séquence d'états. Soulignons que la notion d'asynchronie présentée ici n'a de sens que pour des MMC à topologie gauche-droite et qu'une asynchronie nulle n'est possible qu'entre deux modèles ayant le même nombre d'états.

[1] à l'exception cependant des transitions de probabilité.

Table 1: Corpus DIGIT et LVCSR.

	training		test		
	#spk	#utt	#spk	#utt	#wrd
DIGIT	50	5,490	50	529	4,513
LVCSR	239	17,111	26	207	3,176

2.2. Estimation des paramètres

Les paramètres des MMC multi-flux peuvent être estimés soit indépendamment soit conjointement. Dans la première approche, les paramètres des MMC unimodaux sont estimés indépendamment dans chaque flux avant d'effectuer le produit. Les contraintes de synchronie ne sont alors utilisées que pour le décodage. La deuxième approche consiste à profiter de la représentation sous forme de MMC produit pour estimer conjointement tout les paramètres du modèle. Les contraintes sur la synchronie sont alors utilisées aussi bien pour l'estimation des paramètres que pour le décodage. Dans les deux cas, les coefficients utilisés pour la combinaison des vraisemblances sont déterminés expérimentalement sur un corpus de développement.

Il est à noter que dans l'estimation conjointe des paramètres, les probabilités de transitions dans le modèle produit sont en théorie obtenues à partir des probabilités de transitions des modèles dans chacun des flux. Par la suite, nous avons ignoré ce partage de paramètres et réestimé directement les probabilités de transition du MMC produit. Nous montrons au paragraphe 4.2 que cette approximation n'affecte pas les résultats et simplifie grandement certains aspects pratiques.

3. CONDITIONS EXPÉRIMENTALES

3.1. Modèles et données

Nous considérons tout d'abord une tâche de reconnaissance multi-locuteurs de chiffres connectés. Le système utilise un ensemble de phones en contexte, modélisés dans chaque flux par un MMC gauche-droit à trois états, avec un total de 159 états et 3k Gaussiennes. La deuxième tâche considérée est celle de la reconnaissance de parole lue pour un vocabulaire de 10k mots. Le système utilise également des MMC à trois états pour la modélisation de phones en contexte, avec un total de 3k états et 50k Gaussiennes. Le tableau 1 indique la taille des corpus.

Pour les deux corpus, DIGIT et LVCSR, les enregistrements ont été effectués dans un environnement calme. Un bruit de fond de parole a été ajouté artificiellement pour divers rapport signal à bruit (RSB) sur les corpus d'apprentissage et de test. Le RSB varie de 20 dB pour le corpus d'origine à environ -3 dB. Dans toutes les expériences, les tests ont été faits pour chaque RSB avec des modèles dont les paramètres sont estimés sur des données bruitées au même RSB.

3.2. Représentation des données

Les paramètres acoustiques sont dérivés d'une analyse spectrale par un banc de 24 filtres en échelle MEL. Une analyse linéaire discriminante (LDA) sur 9 trames consécutives est appliquée aux vecteurs spectraux et suivie

d'une transformation linéaire, déterminée selon un critère du maximum de vraisemblance (MLLT), afin de décorréler les coefficients issus de l'analyse discriminante. Les vecteurs acoustiques finaux sont de dimension 60. Les deux transformations, LDA et MLLT, sont estimées pour chaque RSB.

Les paramètres visuels sont calculés à partir d'une région d'intérêt (RI) centrée sur la bouche du locuteur, déterminée automatiquement. Une transformée en cosinus (DCT) est appliquée aux pixels de la RI. Après un suréchantillonnage des coefficients de DCT pour obtenir la même fréquence d'analyse que pour le son, une transformation LDA+MLLT vers un espace de dimension 41 est utilisée sur 15 trames consécutives.

Les paramètres de la LDA, à savoir la taille de la fenêtre de contexte et la dimension de l'espace de projection, ont été déterminés de manière expérimental lors de travaux précédents. Pour comparer l'approche multi-flux avec une approche fusion des représentations, une représentation audiovisuelle des données est obtenue en projetant la concaténation des représentations audio et visuelle définies ci-dessus à l'aide d'une transformation de type LDA+MLLT [6] dans un espace de dimension 60. Cette dernière représentation est définie sous le nom de LDA hiérarchique (HiLDA).

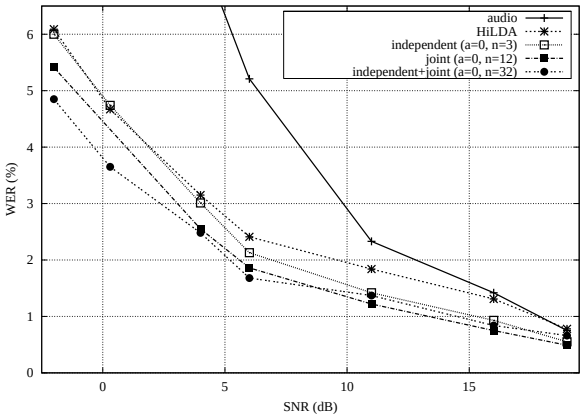
3.3. Décodeur à pile et modèles multi-flux

Toutes les expériences ont été réalisées avec un décodeur à piles multiples [1]. Pour la reconnaissance de chiffres, le décodeur utilise une simple boucle sur les mots du vocabulaire tandis qu'un modèle trigramme est utilisé dans la tâche grand vocabulaire. Des expériences préliminaires à cette étude ont montré que, dans le cas d'une approche par MMC multi-flux asynchrone, les meilleurs résultats sont obtenus lorsque la procédure de sélection rapide de mots candidats pour l'extension des chemins existants s'effectue en utilisant des modèles synchrones plutôt que des modèles asynchrones. Ces derniers sont alors uniquement utilisés pour l'évaluation détaillée des mots candidats sélectionnés. Cette technique accélère le traitement et donne des résultats nettement supérieurs à l'approche utilisant uniquement des modèles asynchrones.

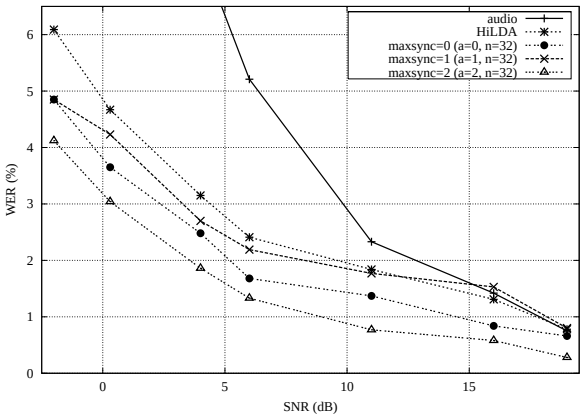
4. RÉSULTATS

4.1. DIGITS

Dans une première expérience, nous comparons les stratégies d'estimation des paramètres discutées au paragraphe 2.2. La figure 2a montre le taux d'erreur (WER) en fonction du RSB pour des modèles synchrones avec 3 stratégies d'estimation: indépendante, conjointe et mixte, c'est-à-dire indépendante suivie d'une réestimation conjointe. La figure montre aussi les résultats pour l'approche HiLDA et pour le système n'utilisant que la représentation acoustique. Les systèmes audiovisuels montrent une robustesse accrue par rapport à l'approche audio et les MMC-MF synchrones améliorent les résultats par rapport à une fusion des représentations. Les meilleurs résultats sont obtenus pour une estimation mixte des paramètres. Des conclusions similaires ont été obtenues avec des MMC multi-flux asynchrones pour lesquels l'avantage de la stratégie mixte d'estimation des paramètres est encore plus marqué.



(*a*) WER pour différentes stratégies d'estimation des paramètres avec des modèles synchrones.



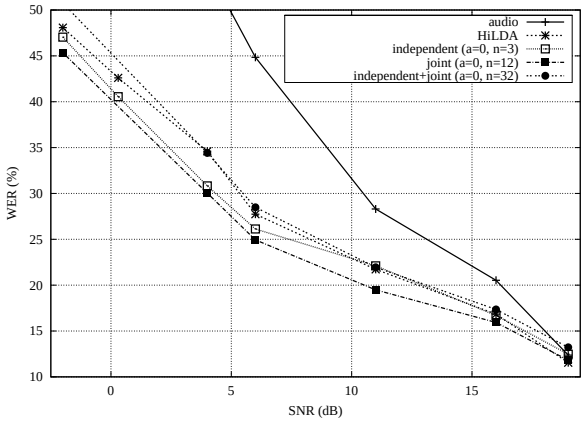
(*b*) WER en fonction du degré d'asynchronie avec une estimation conjointe des paramètres.

Figure 2: Résultats sur la tâche DIGIT.

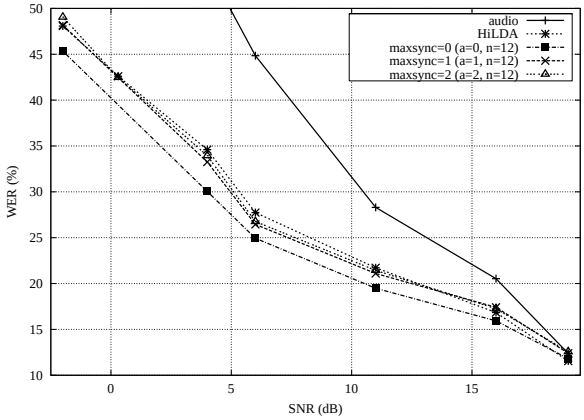
Les résultats figure 2b, obtenus pour des modèles estimés avec une stratégie mixte d'estimation des paramètres, montrent l'influence de l'asynchronie pour $a \in [0, 2]$. A nouveau, les deux systèmes HiLDA et audio sont donnés à titre de comparaison. Les courbes montrent clairement que le taux d'erreur diminue lorsque l'asynchronie entre les frontières de phones est autorisée. Il est intéressant de noter que même à 20 dB, l'intégration asynchrone d'informations visuelles avec l'audio permet d'améliorer les résultats par rapport à l'audio seul avec un WER de 0.28% pour le meilleur système audiovisuel par rapport à 0.75% pour l'audio.

4.2. LVCSR

Les figures 3a et 3b résument les résultats pour la tâche grand vocabulaire. Comme pour les chiffres, les modèles multi-flux synchrones montrent une meilleure robustesse au bruit par rapport aux systèmes audio et HiLDA. Deux observations importantes ressortent de la comparaison des stratégies d'estimation des paramètres. Premièrement, comme observé précédemment, l'estimation jointe donne de meilleurs résultats que l'estimation indépendante des paramètres. En revanche, il est surprenant de noter que la réestimation conjointe de modèles estimés indépendamment au préalable (approche mixte) dégrade sévèrement les résultats, malgré le fait que le modèle produit construit



(*a*) WER pour différentes stratégies d'estimation des paramètres avec des modèles synchrones.



(*b*) WER en fonction du degré d'asynchronie avec une estimation conjointe des paramètres.

Figure 3: Résultat sur la tâche LVCSR.

à partir des modèles estimés indépendamment est meilleur que le modèle utilisé pour initialiser l'estimation conjointe directe. Le même phénomène a été systématiquement observé avec les modèles asynchrones. Une explication possible est que lorsque les modèles audio et visuels sont estimés séparément, les deux modèles convergent vers des solutions très différentes en terme d'alignement. Forcer la synchronie des alignements aux frontières de phones lors de la rééstimation conjointe peut alors conduire à une solution sous-optimale puisque les modèles initiaux ne sont pas adaptés à une telle synchronisation. Nous n'avons pas vérifié cette hypothèse mais, si cela est le cas, cela signifie clairement que la frontière entre phones n'est pas un point de synchronisation pertinent.

Les performances obtenues pour divers degrés d'asynchronie avec des modèles estimés conjointement sont données figure 3b. Le gain observé sur la tâche DIGIT avec l'asynchronie ne se retrouve pas dans ces résultats où, au contraire, l'asynchronie augmente le taux d'erreur. Ceci est probablement dû au fait que des modèles plus complexes sont moins efficaces pour cette tâche où les confusions possibles entre phones sont plus élevées que pour les chiffres. En effet, avec les modèles asynchrones, la complexité de l'espace de recherche augmente considérablement. Cependant, nous n'avons pas d'explication détaillée justifiant ces résultats qui peuvent également être partiellement expliqué par le fait que la frontière entre phones n'est pas un bon point de synchronisation.

Finalement, dans les expériences précédentes, nous n'avons pas considéré le partage des probabilités de transitions. Pour un système asynchrone avec $a = 2$, le WER à 20 dB (resp. 4 dB) est de 12.6% (resp. 33.0%) sans le partage des probabilités de transition et de 12.9% (resp. 33.9%) avec partage. Ce résultat montre bien que l'approximation effectuée n'affecte pas les performances et ne peut donc pas justifier la dégradation due à l'introduction de l'asynchronie.

5. CONCLUSION

Nous avons présenté des expériences sur les MMC multi-flux asynchrones pour la reconnaissance audiovisuelle de la parole. Les résultats montrent les avantages des modèles multi-flux synchrones par rapport à l'approche fusion des représentations. Nous avons observé que les progrès effectué sur une tâche simple de reconnaissance de chiffres ne s'étendaient pas (nécessairement) à une tâche impliquant un vocabulaire plus grand et plus complexe. En particulier, l'utilisation de modèles asynchrones s'est montrée très efficace pour les chiffres alors que l'asynchronie augmente le taux d'erreur sur la tâche grand vocabulaire. Une partie des travaux futurs aura pour but de comprendre ce dernier résultat en regardant les alignements et en construisant des modèles plus simples pour minimiser la complexité de l'espace de recherche lorsqu'on introduit l'asynchronie.

BIBLIOGRAPHIE

[1] L. Bahl, S. De Gennaro, P. Gopalakrishnan, and R.. Mercer. A fast approximate acoustic match for large vocabulary speech recognition. *IEEE Trans. on Speech and Audio Proc.*, 1(1):59–67, January 1993.

[2] T. Chen. Audiovisual speech processing. lip reading and lip synchronization. *IEEE Signal Processing Magazine*, 18:9–21, 2001.

[3] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Trans. on Multimedia*, 2(3):141–151, September 2000.

[4] S. Nakamura. Fusion of audio-visual information for integrated speech processing. In J. Bigun and F. Smeraldi, editors, *Audio- and Video-based Biometric Person Authentication*. Springer-Verlag, 2001.

[5] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio-visual speech recognition. Technical Report Final Workshop Report, Center for Language and Speech Processing, 2000.

[6] G. Potamianos, C. Neti, and J. Luettin. Hierarchical discrimant features for audio-visual LVCSR. In *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Proc.*, 2001.

[7] D. G. Stork and M. E. Hennecke, editors. *Speechreading by Humans and Machines*. Springer, 1996.

[8] M. J. Tomlinson, M. J. Russell, and N. M. Brooke. Integration of audio and visual information to provide highly robust speech recgnition. In *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Proc.*, 1996.