Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Étude comparative des paramètres d'entrée pour la synthèse expressive audiovisuelle de la parole par DNNs

Sara Dahmani[1] Vincent Colotte[1] Slim Ouni[1]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sara.dahmani@loria.fr, vinvent.colotte@loria.fr, slim.ouni@loria.fr

RÉSUMÉ

Dans le passé, les descripteurs contextuels pour la synthèse de la parole acoustique ont été étudiés pour l'entraînement des systèmes basés sur des HMMs. Dans ce travail, nous étudions l'impact de ces facteurs pour la synthèse de la parole audiovisuelle par DNNs. Nous analysons cet impact pour les trois aspects de la parole : la modalité acoustique, la modalité visuelle et les durées des phonèmes. Nous étudions également l'apport d'un entraînement joint et séparé des deux modalités acoustique et visuelle sur la qualité de la parole synthétique générée. Finalement, nous procédons à une validation croisée entre les résultats de la synthèse des différentes émotions. Cette validation croisée, nous a permis de vérifier la capacité des DNNs à apprendre des caractéristiques spécifiques à chaque émotion.

ABSTRACT

Comparative study of input parameters for DNN-based expressive audiovisual speech synthesis

In the past, contextual descriptors for acoustic speech synthesis have been studied for training systems based on HMMs. In this work, we study the impact of these factors for DNN-based audiovisual speech synthesis. We analyze this impact on the three aspects of speech : the acoustic modality, the visual modality and the duration of the phonemes. We also study the contribution of a joint and separate training of the acoustic and visual modalities in the quality of the generated synthetic speech. Finally, we cross-validate the results of the synthesis of the different emotions. This cross validation allowed us to analyze the ability of DNNs to learn characteristics specific to each emotion.

MOTS-CLÉS : Synthèse audiovisuelle expressive, tête parlante expressive, émotion, expression faciale, réseau de neurones profond récurrent à mémoire court-terme et long terme .

KEYWORDS: Expressive audiovisual speech synthesis, Expressive talking head, emotion, facial expression, deep bidirectional long short-term memory neural network (DBLSTM).

1 Introduction

De nos jours, l'animation automatique des têtes parlantes virtuelles expressives est en gain constant d'attention. Elle peut être utilisée dans plusieurs domaines, tel que le domaine des jeux vidéo, des films d'animation ainsi que dans le domaine médical et celui de l'éducation (Sproull *et al.*, 1996; Pandzic *et al.*, 1999; Ostermann & Millen, 2000). L'expressivité dans les systèmes de synthèse de la parole est très demandée puisqu'elle permet d'améliorer l'expérience des utilisateurs et de rendre l'interaction plus naturelle (Eyben *et al.*, 2012; Charfuelan & Steiner, 2013). La synthèse paramétrique statistique de la parole a connu des améliorations ces dernières années, notamment en terme d'intelligibilité (King, 2014), grâce aux techniques paramétriques statistiques allant des HMMs

Volume 1 : Journées d'Études sur la Parole, pages 127–135.
HAL : hal-02798526.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

(Hidden Markov Models) aux réseaux de neurones (Ze *et al.*, 2013; Zen & Senior, 2014), notamment les réseaux BLSTM (Bidirectional Long Short-Term Memory) qui sont capables de prendre en compte les informations passées et futures d'une séquence. Il a été démontré que les BLSTMs donnent de meilleurs résultats de synthèse que les HMMs et les DNNs standards (Fan *et al.*, 2014, 2015; Klimkov *et al.*, 2018). Par ailleurs, le choix des paramètres et la configuration de ces réseaux sont cruciaux pour que la parole synthétique soit naturelle et intelligible. Il faut donc étudier et choisir minutieusement les différents paramètres et architectures impliqués dans l'entraînement des modèles de la parole.

Plusieurs travaux et systèmes se basent sur des descripteurs linguistiques pour des fins de synthèse vocale que ça soit avec des HMMs (Pouget, 2017; Baumann & Schlangen, 2012) ou par DNNs (Wu *et al.*, 2016; Houidhek *et al.*, 2018; Ribeiro *et al.*, 2016). Nous présentons dans cet article une étude bien nécessaire qui regroupe un ensemble d'expériences permettant d'apporter des réponses et des éclaircissements sur le comportement des DNNs face aux données linguistiques et audiovisuelles. Des études similaires ont été menées dans le passé sur des systèmes HMMs (Watts *et al.*, 2010; Le Maguer *et al.*, 2013; Cernak *et al.*, 2013) mais peu d'études se sont intéressées à l'impact des paramètres linguistiques sur un système basé sur les DNNs et encore moins leur impact sur la modélisation de la parole visuelle.

Le Maguer *et al.* (2013) ont étudié l'apport des différents paramètres linguistiques à la qualité de la synthèse du système HTS basé sur des HMMs. Cette étude menée sur un corpus acoustique de langue française a montré que l'utilisation du contexte phonétique améliore la modélisation du spectre de la parole et des durées, et que l'utilisation des informations sur les syllabes améliore la modélisation de la F0. Toutefois, le reste des facteurs contextuels ne semblent pas apporter une amélioration significative à la modélisation acoustique avec HTS. Cernak *et al.* (2013) ont également étudié les facteurs contextuels des données linguistiques pour la synthèse vocale par HMMs pour l'anglais. Cette étude confirme que le contexte syllabique fait partie des facteurs contextuels les plus importants et que le contexte relatif aux mots de la phrase a peu d'importance comme préalablement établie dans l'étude de Yu *et al.* (2010).

Pour la synthèse vocale par DNNs, Ribeiro *et al.* (2016) utilisent différents niveaux de contextes linguistiques pour entraîner un réseau de neurones à propagation vers l'avant (DNN-Feed-Forward ou DNN-FF) pour l'anglais. Les paramètres suprasegmentaux ont été traités par un DNN agissant au niveau des syllabes, et la sortie (sous forme de paramètres acoustiques) de ce dernier a été intégrée en tant qu'entrée supplémentaire à un DNN standard agissant au niveau des *frames*. Cette étude montre que l'ajout d'une représentation pré-entraînée des paramètres suprasegmentaux est bénéfique pour la modélisation acoustique. Par ailleurs, l'ajout des vecteurs de plongement (*embedding*) appris sur des mots ne montre aucune amélioration des performances du DNN. Récemment, Mametani *et al.* (2019) ont présenté une étude des paramètres contextuels appris automatiquement par un système de synthèse *End-to-End* pour l'anglais. Ce genre de système se base sur des DNNs et prend en entrée un texte brut (ou sa représentation phonétique) pour le convertir en paramètres vocaux. Les résultats expérimentaux montrent que le réseau arrive à tirer parti de l'information implicite contenue dans la représentation phonétique du texte comme la réduction des voyelles ou le stress sur les syllabes dans le mot.

Notre étude s'ajoute au travail d'exploration des paramètres d'entrée pour la modélisation de la parole dans la synthèse par DNNs pour la langue française. De plus, dans ce travail nous étudions différents aspects relatifs à la parole, partant des données linguistiques, passant par la modélisation des durées et des données acoustiques et visuelles jusqu'à la modélisation des émotions. Nous effectuons aussi une comparaison objective entre les performances d'un modèle audiovisuel entraîné sur les données

acoustiques et visuelles conjointement puis séparément. Dans le passé, Schabus *et al.* (2013) ont entraîné un système HMMs pour la modélisation audiovisuelle de la parole. Cette étude a montré que les modèles joints offrent une meilleure synchronisation entre les modalités acoustique et visuelle et que la qualité des paramètres acoustiques prédits ne subit pas de dégradation par rapport au modèle acoustique indépendant. Dans une étude similaire, effectuée sur des données audiovisuelles provenant d'une caméra, Filntisis *et al.* (2017) ont déclaré n'avoir trouvé aucune différence significative entre les résultats des deux modèles DNNs (joints et séparés) concernant le réalisme de la vidéo de synthèse. Toutefois, les résultats acoustiques du modèle séparé ont été significativement plus appréciés que ceux du modèle joint. Cette étude s'est basée sur des résultats de tests perceptifs, dans notre étude nous voulons quantifier avec un test objectif l'apport ou la dégradation de la qualité due à l'utilisation d'un modèle joint. De plus, les données visuelles que nous utilisons proviennent d'un système de capture de mouvements et contiennent les informations en 3D. Finalement, nous effectuons une validation croisée sur les résultats obtenus pour les différentes émotions, pour vérifier si les modèles des durées, acoustique et visuel, arrivent à se spécialiser dans la modélisation des différentes émotions.

2 Données utilisées

Dans ce travail nous utilisons le corpus de langue française présenté dans Dahmani *et al.* (2019). Ce corpus a été joué par une actrice semi-professionnelle et contient six émotions plus l'état neutre. Des marqueurs rétro-réflectifs ont été collés sur le visage de l'actrice pour suivre les mouvements de son visage. 2000 phrases ont été enregistrées pour l'état neutre (4h de parole). De ces 2000 phrases, un sous-ensemble de 500 phrases a été sélectionné pour chacune des six émotions basiques : joie, tristesse, colère, surprise, peur et dégoût (entre 55min et 1h 11min de parole pour chaque émotion). Le contenu linguistique est identique pour toutes les émotions et les phrases de ce corpus ont été considérées de telle sorte qu'elles offrent une couverture phonétique maximale. Le corpus neutre couvre 92% des diphones du français et le sous-corpus de 500 phrases en couvre 52%. Les données textuelles, acoustiques et visuelles ont été alignées automatiquement au niveau phonétique.

Nous utilisons un vecteur de 417 paramètres linguistiques composé de :

- 190 paramètres binaires relatifs à la nature du phonème courant et de ses contextes gauches et droits (5x38 paramètres : 36 phonèmes et 2 codes supplémentaires, un pour les pauses et le deuxième pour les silences de début et de fin),
- 195 paramètres binaires relatifs à la catégorie phonétique (voyelle, consonne, nasal, fricatif,...) du phonème courant et de ses contextes gauches et droits (5x39 paramètres),
- 2 paramètres numériques relatifs à la position du phonème courant dans la syllabe courante,
- 7 paramètres numériques relatifs à la syllabe courante précédente et suivante, le nombre de phonèmes qu'elles contiennent, la position de la syllabe courante dans la phrase et dans le mot courants,
- 18 paramètres binaires relatifs à la nature de la voyelle centrale dans la syllabe courante,
- 5 paramètres numériques relatifs aux nombres de syllabes dans le mot courant, précédent et suivant ainsi que la position du mot courant dans la phrase courante.

Ces paramètres linguistiques représentent le vecteur d'entrée pour l'entraînement des trois modèles principaux : des durées, acoustique et visuel.

La durée de chaque phonème est extraite sous forme du nombre de *frames* qu'il couvre en considérant un pas de 5ms entre deux *frames* consécutifs. Pour les paramètres acoustiques, nous avons utilisé le Vocodeur WORLD pour extraire 60 coefficients MFCC (Mel-Frequency Cepstral Coefficients), 5 paramètres BAP (Band-Aperiodicity), la fréquence fondamentale avec une échelle logarithmique (log F0) et leurs paramètres dynamiques (Δ et $\Delta\Delta$) ainsi qu'un paramètre binaire pour préciser la nature

voisée/non-voisée du son dans chaque frame. Ces paramètres ont été extraits des fichiers audio toutes les 5ms. Ils représentent la sortie du DNN qui sera entraîné pour générer des paramètres acoustiques à partir des paramètres linguistiques. Concernant l'aspect visuel, nous nous intéressons dans ce travail uniquement à l'animation de la partie inférieure du visage. Nous sélectionnons sur l'ensemble des données disponibles dans le corpus les 44 points 3D qui couvrent la région des articulateurs (lèvres, joues, mâchoire et menton), soit un vecteur de 132 valeurs. Ces données ont été également présentées avec un écart de 5ms entre deux *frames* consécutifs. Nous avons divisé le corpus en trois sous-ensembles : l'ensemble d'entraînement contenant 80% des données, l'ensemble de validation et celui de test avec 10% de données chacun.

3 Impact du contexte linguistique sur la qualité de la synthèse

Dans cette section, nous testons 4 types différents de paramètres d'entrée pour analyser leurs impacts sur l'apprentissage des modèles de durées, acoustique et visuel :

- 1_cont : Uniquement l'information sur le phonème central ;
- 3_cont : L'information sur le phonème central son contexte gauche et droit immédiats ;
- 5_cont : L'information sur le phonème central ses deux contextes gauches et ses deux droits ;
- 5_cont_p : Même informations que 5_cont en plus des informations sur la position du phonème courant dans la syllabe et la catégorie phonétique des cinq phonèmes du contexte ;
- 5_cont_p_s : Même informations que 5_cont_p en plus des informations sur les syllabes ;
- 5_cont_p_s_m : Même informations que 5_cont_p_s en plus des informations sur les mots ;

Dans cette section nous utilisons uniquement les données du corpus neutre et adoptons deux architectures : une avec des DNN-FF et une autre avec des BLSTMs. Pour ces deux architectures, un DNN à deux couches a été retenu et nous avons essayé plusieurs largeurs de DNNs pour les différents vecteurs d'entrée (256, 512, 1024 et 2048). Les trois modèles ont été entraînés séparément, et l'architecture qui donne le meilleur résultat sur l'ensemble de validation a été retenue pour chaque expérience. Les meilleurs modèles ont été sélectionnés avec la technique du *early stopping*. Les modèles ont été entraînés avec MSE comme fonction de perte. La fonction d'activation des couches cachées est TANH et une fonction d'activation linéaire pour la couche de sortie. Nous avons utilisé l'optimiseur Adam et aucun dropout, BatchNorm ou régularisation spécifique n'a été utilisée.

Le calcul des différentes métriques a été effectué entre les paramètres prédits et ceux provenant du corpus original. Pour le modèle acoustique et visuel, les durées utilisées sont celles provenant du corpus original. Nous affichons la moyenne et les intervalles de confiance pour chaque métrique.

	DNN-FF					
	1_cont [256, 256]	3_cont [256, 256]	5_cont [512, 512]	5_cont_p [512, 512]	5_cont_p_s [512, 512]	5_cont_p_s_m [512, 512]
RMSE (f/p)	8.672 (±0.018)	5.888 (±0.007)	5.532 (±0.006)	5.435 (±0.009)	5.259 (±0.008)	**5.256** (±0.007)
Corrélation	0.389 (±0.002)	0.781 (±0.0008)	0.811 (±0.0004)	0.822 (±0.0007)	0.827 (±0.0007)	**0.827** (±0.0007)
	BLSTM					
	1_cont [256, 256]	3_cont [256, 256]	5_cont [512, 512]	5_cont_p [512, 512]	5_cont_p_s [512, 512]	5_cont_p_s_m [512, 512]
RMSE (f/p)	7.237 (±0.011)	5.418 (±0.006)	5.413 (±0.006)	5.411 (±0.006)	5.301 (±0.007)	**5.247** (±0.006)
Corrélation	0.639 (±0.002)	0.821 (±0.0007)	0.824 (±0.00006)	0.826 (±0.0006)	0.827 (±0.0006)	**0.827** (±0.0006)

TABLE 1 – *Les résultats du RMSE en frames/phonème et de la corrélation de Pearson sur l'ensemble de test générés par le modèle de durées en variant les paramètres linguistiques lors de l'entraînement avec une architecture de type DNN-FF et BLSTM.*

Dans les tableaux 1, 2 et 3, il est très intéressant de constater qu'en utilisant une architecture DNN-FF, l'ajout de toutes les informations contextuelles améliore la qualité de la synthèse pour les trois aspects de la parole, bien que l'écart soit extrêmement serré entre les résultats avec et sans informations relatives aux mots. Toutefois, pour le réseau de type BLSTM les trois modèles n'ont pas tous le

	DNN-FF					
	1_cont [256, 256]	3_cont [512, 512]	5_cont [1024, 1024]	5_cont_p [1024, 1024]	5_cont_p_s [1024, 1024]	5_cont_p_s_m [1024, 1024]
MCD (dB)	6.653 (±0.026)	6.136 (±0.025)	6.132 (±0.024)	5.910 (±0.024)	5.901 (±0.024)	**5.900** (±0.024)
BAPD (dB)	0.327 (±0.004)	0.295 (±0.004)	0.292 (±0.004)	0.295 (±0.003)	0.288 (±0.003)	**0.287** (±0.003)
F0-RMSE (Hz)	35.226 (±1.105)	32.447 (±1.132)	31.334 (±1.106)	31.360 (±0.757)	30.648 (±0.733)	**30.555** (±0.743)
F0-Corrélation	0.341 (±0.021)	0.481 (±0.018)	0.529 (±0.017)	0.526 (±0.016)	0.557 (±0.016)	**0.563** (±0.015)
V/N-V (%)	14.250 (±0.424)	13.195 (±0.539)	13.090 (±0.553)	12.877 (±0.371)	12.872 (±0.375)	**12.848** (±0.371)
	BLSTM					
	1_cont [256, 256]	3_cont [512, 512]	5_cont [1024, 1024]	5_cont_p [1024, 1024]	5_cont_p_s [1024, 1024]	5_cont_p_s_m [1024, 1024]
MCD (dB)	5.304 (±0.029)	**5.099** (±0.023)	5.152 (±0.023)	5.103 (±0.025)	5.106 (±0.026)	5.146 (±0.024)
BAPD (dB)	0.282 (±0.004)	**0.242** (±0.003)	0.245 (±0.003)	0.242 (±0.003)	0.247 (±0.003)	0.247 (±0.003)
F0-RMSE (Hz)	32.580 (±0.818)	**27.934** (±0.622)	29.010 (±0.624)	28.460 (±0.690)	28.201 (±0.648)	28.207 (±0.865)
F0-Corrélation	0.471 (±0.016)	**0.640** (±0.013)	0.620 (±0.013)	0.628 (±0.014)	0.639 (±0.013)	0.637 (±0.014)
V/N-V (%)	10.736 (±0.369)	**8.348** (±0.262)	8.822 (±0.257)	8.571 (±0.293)	8.566 (±0.303)	8.755 (±0.292)

TABLE 2 – *Les résultats sur l'ensemble de test générés par le modèle acoustique en variant les paramètres linguistiques lors de l'entraînement avec une architecture de type DNN-FF et BLSTM.*

	DNN-FF					
	1_cont [256, 256]	3_cont [256, 256]	5_cont [512, 512]	5_cont_p [1024, 1024]	5_cont_p_s [1024, 1024]	5_cont_p_s_m [1024, 1024]
RMSE (mm)	1.760 (±0.031)	1.458 (±0.029)	1.429 (±0.030)	1.427 (±0.030)	1.424 (±0.029)	**1.423** (±0.029)
Corrélation	0.574 (±0.007)	0.763 (±0.005)	0.778 (±0.006)	0.778 (±0.005)	0.779 (±0.005)	**0.779** (±0.005)
	BLSTM					
	1_cont [256, 256]	3_cont [256, 256]	5_cont [512, 512]	5_cont_p [1024, 1024]	5_cont_p_s [1024, 1024]	5_cont_p_s_m [1024, 1024]
RMSE (mm)	1.327 (±0.030)	**1.316** (±0.029)	1.328 (±0.031)	1.330 (±0.030)	1.331 (±0.031)	1.332 (±0.031)
Corrélation	0.823 (±0.005)	**0.828** (±0.005)	0.822 (±0.005)	0.822(±0.005)	0.821 (±0.005)	0.821 (±0.005)

TABLE 3 – *Les résultats sur l'ensemble de test générés par le modèle visuel en variant les paramètres linguistiques lors de l'entraînement avec une architecture de type DNN-FF et BLSTM.*

même comportement face aux informations linguistiques. Pour le modèle des durées, et de manière similaire au DNN-FF, l'ajout de l'ensemble des informations linguistiques améliore la prédiction des durées. Concernant les modèles acoustique et visuel, le réseau BLSTM atteint la meilleure qualité de synthèse avec les contextes gauche et droit immédiats uniquement. Cela peut s'expliquer par la capacité des BLSTMs à accéder automatiquement aux contextes passés et futurs de la frame courante, contrairement aux DNN-FF qui nécessitent que cette information soit explicitement donnée en entrée.

Nous remarquons qu'en utilisant le réseau BLSTM, pour le modèle acoustique, les informations sur la position et la catégorie des phonèmes ainsi que les informations sur les syllabes améliorent la modélisation de la F0, alors que l'ajout des informations relatives aux mots a un impact presque nul sur les mesures objectives. Ces constatations confirment les résultats des études précédentes (Le Maguer *et al.*, 2013; Cernak *et al.*, 2013; Yu *et al.*, 2010). Par ailleurs, force est de constater que, pour le modèle visuel, l'ajout des informations contextuelles autres que les contextes gauche et droit immédiats est néfaste pour l'apprentissage. Ce comportement peut être expliqué par la réduction du nombre d'exemples d'apprentissage avec l'augmentation du nombre de combinaisons possibles dans le vecteur d'entrée. En réalité, l'ajout de plus de contraintes contextuelles divise les données en classes de plus en plus petites, et réduit de ce fait le nombre d'exemples d'apprentissage de chaque classe. Pour le modèle des durées, ce comportement ne semble pas se produire, nous pensons que cela vient du fait que le réseau doit prédire un seul et unique paramètre, qui est une tâche plus simple et qui nécessite donc moins d'exemples d'apprentissage.

4 Entraînement joint et séparé des modèles acoustique et visuel

Dans cette section nous étudions l'apport éventuel d'un entraînement joint des modalités acoustique et visuelle sur la qualité de la synthèse audiovisuelle. Nous incluons les six catégories d'émotions dans le processus d'apprentissage et nous utilisons 3_cont comme informations linguistiques. Le

vecteur de sortie pour le modèle joint est le résultat de la concaténation des paramètres acoustiques et visuels.

Le calcul des différentes métriques a été effectué entre les paramètres prédits et ceux provenant du corpus original. Pour le modèle acoustique et visuel, les durées utilisées sont celles provenant du corpus original.

	Modèles Séparés							Modèles joints [2048, 2048]						
	Neu	Joi	Tri	Col	Sur	Peu	Dég	Neu	Joi	Tri	Col	Sur	Peu	Dég
	Acoustique [1024, 1024]							Acoustique						
MCD (dB)	**4.863**	**5.738**	**5.288**	**5.262**	**5.699**	**5.226**	**5.431**	5.305	6.135	5.740	5.691	6.157	5.669	5.844
BAPD (dB)	**0.224**	**0.312**	**0.269**	**0.268**	**0.287**	**0.231**	**0.256**	0.265	0.359	0.304	0.322	0.335	0.269	0.304
F0-RMSE (Hz)	**26.172**	**46.723**	**32.074**	**39.514**	**32.203**	**40.617**	**35.972**	32.203	47.617	37.972	45.094	45.676	46.201	44.003
F0-Corrélation	**0.687**	**0.631**	**0.518**	**0.524**	**0.702**	**0.627**	**0.535**	0.683	0.627	0.514	0.513	0.683	0.488	0.518
V/N-V (%)	**6.900**	**10.167**	**7.692**	**8.082**	**9.874**	**7.711**	**9.137**	7.851	11.879	8.955	9.587	11.864	8.814	10.560
	Visuel [1024, 1024]							Visuel						
RMSE (mm)	**1.304**	**1.572**	**1.317**	**1.466**	**1.482**	**1.424**	**2.124**	1.309	1.581	1.320	1.475	1.504	1.429	2.132
Corrélation	**0.833**	**0.777**	**0.792**	**0.810**	**0.807**	**0.826**	**0.696**	0.829	0.776	0.790	0.808	0.803	0.825	0.689

TABLE 4 – *Les résultats des paramètres acoustiques et visuels sur l'ensemble de test générés en entraînant le DNN avec les données acoustiques et visuelles séparément puis conjointement.*

Le tableau 4 montre les résultats obtenus avec les deux modèles. Nous remarquons que l'entraînement joint des deux modalités dégrade toutes les mesures objectives, que ça soit pour la modalité acoustique ou visuelle. En effectuant une écoute informelle nous avons constaté plus de distorsion et un son légèrement étouffé dans les résultats acoustiques du modèle joint, mais pour les résultats visuels, nous n'avons constaté aucune différence humainement perceptible. Ce résultat rejoint celui de Filntisis *et al.* (2017) qui a montré via des tests perceptifs que les résultats des modèles séparés sont considérés comme légèrement plus réalistes, mais qu'aucune différence d'ordre significatif n'a été trouvée entre les résultats audiovisuels des deux modèles. Cependant, les résultats acoustiques du modèle séparé ont été considérés comme significativement plus réalistes que ceux générés par le modèle joint.

5 Validation croisée des résultats de la synthèse expressive

Dans cette expérience nous utilisons des modèles acoustique et visuel séparés avec 3_cont comme vecteur d'entrée, et 5_cont_p_s_m comme entrée du modèle des durées. Sachant que dans une étude précédente (Dahmani *et al.*, 2019), utilisant le même corpus et une architecture neuronale similaire (BLSTM), il a été montré, via des tests perceptifs que les émotions synthétiques sont correctement reconnues (sauf la peur et la tristesse). Dans ce travail nous souhaitons vérifier, via une étude objective la capacité des modèles à apprendre des caractéristiques spécifiques à chaque émotion. Pour ce faire nous procédons à une validation croisée.

Dans cette expérience, nous évaluons la capacité de nos modèles à modéliser les durées et les modalités acoustique et visuelle, toutefois la prononciation des phrases peut changer d'une émotion à l'autre (plus ou moins de pauses, suppression/ajout de voyelles). Ce dernier point n'est pas étudié dans ce travail. Pour le modèle des durées, nous utilisons les informations linguistiques de l'ensemble de test d'une émotion cible pour générer les durées de toutes les autres émotions et nous avons calculé les mesures de toutes les autres émotions par rapport aux données originales de l'émotion cible. Pour les modèles acoustique et visuel, nous avons considéré les données linguistiques ainsi que les durées des données originales de l'ensemble de test de l'émotion cible. En utilisant ces informations, nous générons les paramètres acoustiques et visuels correspondants à chaque émotion et calculons les différentes mesures. Les résultats relatifs à chaque émotion traitée sont représentés dans les lignes des tableaux 5, 6 et 7. Les résultats affichés dans ces trois tableaux montrent que les trois modèles arrivent à se spécialiser dans la modélisation des différentes émotions. Pour le modèle des durées, le dégoût semble être très différent des autres émotions. Cela peut s'expliquer par les durées des

Durées								
		Neutre	Joie	Tristesse	colère	surprise	Peur	Dégoût
Neutre	RMSE (f/p)	**5.289**	6.110	6.001	5.917	5.652	6.378	13.280
	Corrélation	**0.831**	0.799	0.804	0.786	0.803	0.806	0.779
Joie	RMSE (f/p)	7.346	**7.136**	7.385	7.272	7.206	7.708	14.703
	Corrélation	0.752	**0.774**	0.756	0.760	0.769	0.751	0.720
Tristesse	RMSE (f/p)	6.886	6.881	**6.606**	7.118	7.176	6.926	13.856
	Corrélation	0.770	0.765	**0.777**	0.755	0.754	0.770	0.747
colère	RMSE (f/p)	6.879	7.130	7.222	**6.463**	7.597	6.578	15.195
	Corrélation	0.720	0.737	0.728	**0.758**	0.729	0.744	0.686
surprise	RMSE (f/p)	6.394	6.905	7.134	6.471	**6.006**	7.532	14.582
	Corrélation	0.756	0.763	0.741	0.753	**0.781**	0.749	0.708
Peur	RMSE (f/p)	7.573	7.468	7.287	7.760	7.789	**7.174**	13.578
	Corrélation	0.767	0.758	0.766	0.756	0.763	**0.781**	0.753
Dégoût	RMSE (f/p)	13.614	15.709	13.669	15.162	14.361	13.146	**9.311**
	Corrélation	0.728	0.716	0.723	0.693	0.712	0.721	**0.741**

TABLE 5 – *Les résultats du RMSE en frames/phonème et de la corrélation de Pearson pour la validation croisée sur les résultats de prédiction des durées des données expressives de l'ensemble de test.*

Visuel									
		Neutre	Joie	Tristesse	colère	surprise	Peur	Dégoût	Statique
Neutre	RMSE (mm)	**1.304**	2.392	1.635	2.464	1.945	2.245	2.377	2.170
	Corrélation	**0.833**	0.77	0.801	0.769	0.782	0.805	0.739	—
Joie	RMSE (mm)	2.500	**1.572**	2.125	2.703	2.605	2.814	2.732	3.217
	Corrélation	0.727	**0.777**	0.734	0.722	0.736	0.734	0.712	—
Tristesse	RMSE (mm)	1.655	2.092	**1.317**	2.378	2.241	2.221	2.325	2.364
	Corrélation	0.775	0.753	**0.792**	0.723	0.727	0.773	0.713	—
colère	RMSE (mm)	2.604	2.564	2.439	**1.466**	2.100	1.688	3.124	3.308
	Corrélation	0.732	0.735	0.714	**0.810**	0.783	0.774	0.716	—
surprise	RMSE (mm)	1.984	2.537	2.271	2.046	**1.482**	1.980	2.614	2.817
	Corrélation	0.750	0.744	0.723	0.785	**0.807**	0.771	0.727	—
Peur	RMSE (mm)	2.255	2.778	2.239	1.715	1.883	**1.424**	3.041	3.055
	Corrélation	0.794	0.772	0.791	0.795	0.790	**0.826**	0.748	—
Dégoût	RMSE (mm)	2.823	3.160	2.822	3.414	3.063	3.460	**2.124**	3.530
	Corrélation	0.651	0.647	0.641	0.644	0.651	0.649	**0.696**	—

TABLE 6 – *Les résultats de validation croisée sur les résultats de prédiction des trajectoires visuelles des données expressives de l'ensemble de test. Statique représente un visage à l'état neutre avec une bouche constamment fermée.*

émotions dans le corpus utilisé. En fait, cette émotion a été jouée avec un débit remarquablement lent. La durée du corpus du dégoût (1h 53min) représente environ le double des durées des autres émotions (entre 55min et 1h 11min). Les résultats visuels nous permettent de voir certaines ressemblances entre quelques émotions, notamment entre l'état neutre et la tristesse et entre la colère et la peur. En ce qui concerne le modèle acoustique, nous remarquons qu'il y a également une ressemblance entre l'état neutre et la tristesse puis entre la colère et le dégoût, de plus la joie et la surprise sont les émotions avec le plus grand écart de F0 par rapport au neutre et aux autres émotions.

6 Conclusion

Dans cet article, nous avons effectué une étude bien nécessaire sur la synthèse audiovisuelle expressive de la parole, afin de donner des éclaircissements sur l'apport de certains paramètres sur les résultats générés. Pour atteindre cet objectif, nous avons adopté différentes architectures neuronales pour entraîner trois modèles : le modèle des durées, le modèle acoustique et le modèle visuel. Nous avons réalisé une comparaison directe entre ces architectures en variant les paramètres linguistiques utilisés. Les résultats obtenus montrent que bien que toutes les informations linguistiques soient bénéfiques pour le modèle des durées, pour le modèle acoustique, uniquement les informations sur le contexte gauche et droit immédiats ainsi que le contexte syllabique améliore la prédiction. Toutefois pour le modèle visuel les informations autres que le contexte gauche et droit immédiats semblent être nuisibles pour l'apprentissage. Nous avons également comparé la qualité de la synthèse des modèles

		Acoustique						
		Neutre	Joie	Tristesse	colère	surprise	Peur	Dégoût
Neutre	MCD (dB)	**4.863**	6.409	5.390	5.784	6.548	5.327	5.539
	BAPD (dB)	**0.224**	0.304	0.243	0.268	0.263	0.229	0.232
	F0-RMSE (Hz)	**26.172**	97.521	33.810	42.063	108.220	35.640	38.839
	F0-Corrélation	**0.687**	0.610	0.598	0.546	0.404	0.558	0.604
	V/N-V (%)	**6.900**	7.565	7.594	7.512	7.612	7.154	7.486
Joie	MCD (dB)	7.010	**5.738**	6.696	6.417	6.132	7.227	7.045
	BAPD (dB)	0.367	**0.312**	0.347	0.330	0.334	0.377	0.371
	F0-RMSE (Hz)	103.444	**46.723**	85.568	97.438	58.942	113.167	109.806
	F0-Corrélation	0.586	**0.631**	0.552	0.547	0.455	0.526	0.540
	V/N-V (%)	11.015	**10.167**	10.792	10.751	10.547	10.812	10.908
Tristesse	MCD (dB)	5.688	6.442	**5.288**	5.825	6.642	5.660	5.817
	BAPD (dB)	0.271	0.301	**0.269**	0.284	0.284	0.271	0.271
	F0-RMSE (Hz)	33.943	82.658	**32.074**	39.353	96.357	47.107	44.815
	F0-Corrélation	0.503	0.476	**0.518**	0.496	0.284	0.509	0.514
	V/N-V (%)	8.107	8.246	**7.692**	8.167	8.258	7.984	8.023
colère	MCD (dB)	6.177	6.069	5.793	**5.262**	6.171	6.039	6.040
	BAPD (dB)	0.303	0.287	0.290	**0.268**	0.291	0.303	0.304
	F0-RMSE (Hz)	43.043	89.370	41.357	**39.514**	97.935	44.919	43.601
	F0-Corrélation	0.440	0.454	0.497	**0.524**	0.357	0.505	0.491
	V/N-V (%)	8.705	8.615	8.515	**8.082**	8.807	8.347	8.495
surprise	MCD (dB)	6.806	5.916	6.616	6.277	**5.699**	6.951	6.911
	BAPD (dB)	0.305	0.30	0.302	0.301	**0.287**	0.311	0.311
	F0-RMSE (Hz)	102.248	51.066	86.765	97.706	**32.203**	112.539	109.363
	F0-Corrélation	0.449	0.564	0.394	0.444	**0.702**	0.382	0.391
	V/N-V (%)	10.176	9.769	9.967	10.078	**9.874**	10.007	10.105
Peur	MCD (dB)	5.730	7.015	5.729	6.097	7.075	**5.252**	5.226
	BAPD (dB)	0.246	0.334	0.262	0.297	0.286	**0.234**	0.231
	F0-RMSE (Hz)	37.586	116.012	48.980	41.281	128.206	**32.505**	35.090
	F0-Corrélation	0.435	0.404	0.487	0.477	0.242	**0.494**	0.627
	V/N-V (%)	7.951	8.254	8.307	8.258	8.352	**7.649**	7.711
Dégoût	MCD (dB)	5.995	7.124	5.949	6.250	7.269	5.641	**5.431**
	BAPD (dB)	0.272	0.338	0.279	0.328	0.296	0.265	**0.256**
	F0-RMSE (Hz)	38.890	117.422	46.779	42.528	132.273	36.477	**35.972**
	F0-Corrélation	0.473	0.439	0.512	0.502	0.299	0.516	**0.535**
	V/N-V (%)	9.350	9.840	9.446	9.837	9.828	9.245	**9.137**

TABLE 7 – *Les résultats de validation croisée sur les résultats de prédiction des paramètres acoustiques des données expressives de l'ensemble de test.*

acoustique et visuel entraînés séparément puis conjointement. Nous avons trouvé que les modèles entraînés séparément atteignent une meilleure précision de reconstruction lors de la comparaison via des tests objectifs. Finalement, les résultats objectifs de la validation-croisée effectuée sur les différentes émotions, montrent que les trois modèles arrivent à se spécialiser aux différentes émotions. Ces résultats nous ont aussi permis de constater des similarités et des différences entre certaines émotions. Ces résultats viennent pour compléter une étude perceptive effectuée précédemment par Dahmani *et al.* (2019) sur le même corpus. Sachant que les résultats objectifs ne reflètent pas toujours la perception humaine, nous comptons compléter, dans nos prochains travaux, cette étude par des tests perceptifs.

Nous souhaitons que cet ensemble d'expériences apportera plus de clarté sur le comportement des DNNs face aux différentes données linguistiques, des durées et audiovisuelles, et que ça facilitera, pour les autres chercheurs, le choix de l'architecture neuronale la plus adaptée pour la synthèse audiovisuelle expressive de la parole.

Remerciements

Nous remercions la plateforme Grid'5000 de nous avoir fourni des ressources GPU pour entraîner nos modèles (Balouek *et al.*, 2012).

Références

BALOUEK D., AMARIE *et al.* (2012). Adding virtualization capabilities to the Grid'5000 testbed. In *International Conference on Cloud Computing and Services Science* : Springer.

BAUMANN T. & SCHLANGEN D. (2012). Inpro_iss : A component for just-in-time incremental speech synthesis. In *Proceedings of the ACL 2012 System Demonstrations*.

CERNAK M., MOTLICEK P. & GARNER P. N. (2013). On the (un) importance of the contextual factors in hmm-based speech synthesis and coding. In *ICASSP 2013* : IEEE.

CHARFUELAN M. & STEINER I. (2013). Expressive speech synthesis in MARY TTS using audiobook data and emotionML. In *INTERSPEECH*.

DAHMANI S., COLOTTE V., GIRARD V. & OUNI S. (2019). Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis. *Proc. Interspeech 2019*.

EYBEN F., BUCHHOLZ *et al.* (2012). Unsupervised clustering of emotion and voice styles for expressive TTS. In *ICASSP 2012* : IEEE.

FAN B. *et al.* (2015). Photo-real talking head with deep bidirectional lstm. In *ICASSP*.

FAN Y. *et al.* (2014). Tts synthesis with bidirectional lstm based recurrent neural networks. In *INTERSPEECH*.

FILNTISIS P. P., KATSAMANIS A., TSIAKOULIS P. & MARAGOS P. (2017). Video-realistic expressive audio-visual speech synthesis for the greek language. *Speech Communication*, **95**.

HOUIDHEK A., COLOTTE V., MNASRI Z. & JOUVET D. (2018). Dnn-based speech synthesis for arabic : modelling and evaluation. In *International Conference on Statistical Language and Speech*.

KING S. (2014). Measuring a decade of progress in text-to-speech. *Loquens*, **1**(1).

KLIMKOV V., MOINET A. *et al.* (2018). Parameter generation algorithms for text-to-speech synthesis with recurrent neural networks. In *SLT Workshop 2018* : IEEE.

LE MAGUER S. L., BARBOT N. & BOEFFARD O. (2013). Evaluation of contextual descriptors for hmm-based speech synthesis in french. In *Eighth ISCA Workshop on Speech Synthesis*.

MAMETANI K., KATO T. & YAMAMOTO S. (2019). Investigating context features hidden in end-to-end tts. In *ICASSP 2019* : IEEE.

OSTERMANN J. & MILLEN D. (2000). Talking heads and synthetic speech : An architecture for supporting electronic commerce. In *2000 ICME2000* : IEEE.

PANDZIC I. S., OSTERMANN J. & MILLEN D. (1999). User evaluation : Synthetic talking faces for interactive services. *The visual computer*, **15**(7-8).

POUGET M. (2017). *Synthèse incrémentale de la parole à partir du texte*. Thèse de doctorat.

RIBEIRO M. S., WATTS O. & YAMAGISHI J. (2016). Syllable-level representations of suprasegmental features for dnn-based text-to-speech synthesis. In *INTERSPEECH*.

SCHABUS D., PUCHER M. & HOFER G. (2013). Joint audiovisual hidden semi-markov model-based speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, **8**(2).

SPROULL L., *et al.* (1996). When the interface is a face. *Human-Computer Interaction*.

WATTS O. *et al.* (2010). The role of higher-level linguistic features in hmm-based speech synthesis.

WU Z. *et al.* (2016). Merlin : An open source neural network speech synthesis system. In *SSW*.

YU K. *et al.* (2010). Word-level emphasis modelling in hmm-based speech synthesis. In *ICASSP*.

ZE H. *et al.* (2013). Statistical parametric speech synthesis using deep neural networks. In *ICASSP*.

ZEN H. & SENIOR A. (2014). Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *ICASSP*.