UNIVERSITÉ PARIS-SUD
Ecole Doctorale d'Informatique
Spécialité: Informatique

# Text-To-Speech à base de HMM (Hidden Markov Model) pour le vietnamien: modélisation de la segmentation prosodique, la conception du corpus, la conception du système, et l'évaluation perceptive

Doctorante: NGUYEN Thi Thu Trang, trangntt.it@gmail.com
International join PhD,
Université Paris-Sud 11 and Hanoi University of Science and Technology

Directeur de researche: Christophe D'ALESSANDRO, LIMSI-CNRS
Thi Ngoc Yen PHAM, MICA-CNRS
Do Dat TRAN, MICA-CNRS

L'objectif de cette thèse est de concevoir et de construire un système de synthèse de la parole a partir du texte de haute qualité basé sur le modèle de Markov caché (HMM) pour la langue Vietnamienne - une langue tonale. Le système est appelé VTED (Vietnamese TExt-to-speech Development system). Les contributions principales de ce travail peuvent être récapitulées comme suit: (i) proposition d'une nouvelle unité de la parole (unité acoustique) - tonophone – pour la synthèse de la parole vietnamienne, (ii) conception, enregistrement, et prétraitement d'un nouveau corpus, VDTS, qui couvre contexte phonémique ainsi que contexte tonale, pour le TTS vietnamienne, (iii) proposition d'un modèle de phrasé prosodique en utilisant les informations syntaxique automatique pour le TTS vietnamienne, (iv) conception et construction d'un système de synthèse de la parole complet en langue vietnamienne basé sur le HMM composant la partie de traitement du langage naturel, et (v) évaluation du système TTS avec plusieurs tests de perception concernant les tons lexicaux.

## 1. Proposer une nouvelle unité de la parole - tonophone

Les syllabes vietnamiennes ont une structure hiérarchique. Le premier niveau de la syllabe est constitué de deux parties principales: une consonne initiale et une rime. Le ton est une partie non-linéaire ou suprasegmentale d'une syllabe, et principalement adhère à la rime. Les tons apparaissent simultanément avec des éléments constituants de la rime incluant un son médian, un son de noyau et un son final. Le son noyau et le ton sont obligatoires tandis que d'autres sont facultatifs. Bien qu'il y ait six tons lexicaux dans le système d'écriture, la langue vietnamienne a un paradigme de six-ton pour les syllabes sonrée-finale: le ton plat 1, le ton descendant 2, le ton interrogatif 3, le ton brisé

4, le ton montant 5a, et le ton grave 6a ; plus un paradigme de deux tons pour syllabes obstruantes-finale: le ton montant 5b et le ton grave 6b.

En se fondant sur l'étude de la littérature, en raison de la grande importance des tons lexicaux, un « tonophone » - un allophone dans le contexte tonal - a été proposé comme une nouvelle unité acoustique pour notre travail. Pour construire l'ensemble de tonophones du système, le ton lexical a été pris en compte et adhéré à tous les allophones dans la rime, et la consonne initiale a maintenu sa forme sans aucune information de la tonalité. En conséquence, un ensemble de tonophones avec 207 tonophones a été construit à partir de 48 allophones vietnamiens. Cet ensemble comprend: (i) 19 consonnes initiales sans information de tonalité, (ii) son médian et 16 sons noyaux adhérant à huit tons, (iii) sons finaux (consonnes non aspirées occlusives) adhérant à deux tons 5b, 6b, et (iv) d'autres consonnes finales adhérant à six tons 1-4, 5a, 6a. Un ensemble de tonophones acoustiques-phonétiques vietnamien a également été construit pour (i) « HMM clustering » en utilisant des arbres de décision phonétiques, et (ii) l'étiquetage automatique, incluant segmentation automatique et force d'alignement du corpus de parole avec les transcriptions orthographiques. Basé sur la revue de la littérature, les attributs phonétiques principaux ont été attachés aux consonnes et aux voyelles pour cet ensemble d'unités acoustique-phonétiques, tels que le lieu ou l'articulation ou le mode d'articulation pour les consonnes, la position de la langue ou de la hauteur pour les voyelles.

Des règles graphème à phonème ont été développées pour la transcription des consonnes et des voyelles/diphtongues vietnamiennes. Beaucoup de graphèmes peuvent être directement convertis en tonophones sans aucune ambiguïté, comme "b-"- [á], "ch-, tr-" - [tC], "-m"- [m], "ê"-[e]. Des règles bien définies ont été trouvées pour les cas/ variantes plus complexes. Par exemple, pour le graphème «a», s'il est suivi par "nh" ou "ch", le phonème est [e]; s'il est suivi par "u" ou "y", le phonème est [a]; autrement, le phonème est [a]. Les règles complètes de G2P ont été utilisées à la fois pour transcrire le texte brut pour la conception du corpus et pour la construction du module de conversion G2P de notre système de TTS.

PRO-SYLDIC, un e-dictionnaire de prononciation des syllabes vietnamiennes a été construit pour filtrer syllabes prononçables dans la normalisation de texte ainsi que la transcription des textes. Paires de l'orthographe syllabe et la transcription dans le dictionnaire ont été automatiquement générés principalement basée sur (i) les règles de G2P, (ii) des règles orthographiques syllabes, et (iii) la liste des rimes. Une table de 170 rimes vietnamiens existent ainsi que prononçables dans le langage a été conçu. La raison de maintenir toutes les rimes prononçables est que l'entrée d'un système de TTS vietnamien peut inclure de nombreux de mots emprunts qui comprend syllabes inexistants mais prononçables, ainsi que des mots nouveaux utilisés par les jeunes ou les utilisateurs d'Internet. Le PRO-SYLDIC a été construit en combinant 19 consonnes initiales avec des 772 rimes portant tons, soit un total de 15,440 syllabes prononçables (d'orthographe).

## 2. Concevoir et construire un nouveau corpus

Un nouveau corpus était toujours une grande préoccupation durant le processus de développement d'un système TTS. Une méthode pour construire un corpus phonétiquement riches et balances en termes de contextes phonémiques et tonals à partir d'un grand texte a été présenté. Les tâches suivante a été appliquée sur la conception de nouveaux corpus: (i) collection du texte brut, (ii) prétraitement du texte brut, (iii) conception du corpus, (iv) l'enregistrement du corpus, et (v) prétraitement du corpus. Bien que la plupart de ces tâches aient été programmés pour effectuer automatiquement un grand nombre de tâches manuelles avait encore beaucoup à faire. La raison était qu'il y avait beaucoup de cas exceptionnels, ainsi que les erreurs humaines.

### 2.1. Construire un grand corpus textuel brut phonétiquement riches et balances

Puisque le corpus de texte brut a été considérée comme étant représentée pour la langue en termes de distribution phonétique dans la conception de corpus, il est collecté à partir de différentes ressources (telles que des journaux, des e-histoires, des exemples de l'e-dictionnaire vietnamien, les ressources existantes, et la conception spéciale). Il a besoin d'un pré-processus afin de l'adapter pour le processus de conception qui contient cinq tâches principales: (i) Segmentation de phrase, (ii) Tokenisation, (iii) nettoyage du texte, (iv) Normalisation du texte, et (v) transcription du texte. Textes ont d'abord été segmentés en phrases, puis tokenizés en syllabes ou les mots non standard (NSWs - qui ne peuvent être directement transcrits aux phonèmes, par exemple, numéros, dates, abréviations, monnaie). Sentences ayant plus de 70 syllabes ou contenant des caractères imprononçables (par exemple, celles de contrôle) ont été enlevés. La prochaine tâche pour les phrases "propres" est la normalisation de texte, dans lequel NSWs ont été traitées et ensuite élargis en syllabes prononçables. Le texte normalisé a finalement été transcrit en tonophones afin d'avoir une entrée appropriée pour les étapes suivantes. Le corpus de texte brut inclus 323,934 phrases propres, et plus de 10 milliards de syllabes.

La langue vietnamienne est une langue tonale, donc notre travail vise à concevoir un corpus phonétiquement riches et balances à la fois le contexte phonémique et le contexte tonal. Alors, deux unités acoustiques proposées sont utilisées pour la conception de corpus pour systèmes de TTS en langue vietnamienne, elles sont: (i) un "tonophone": un phoneme concernant le ton lexical de la syllabe, et (ii) un "di-tonophone": une paire adjacente de "tonophones". L'ensemble di-tonophone a été construit en utilisant un dictionnaire composant les syllabes significatives (théoriques), et en utilisant le corpus de texte brut (réel). La distribution phonétique du corpus de texte brut, qui a été considéré comme une référence pour la conception de corpus, a été calculé pour les différentes unités de la parole, y compris les deux nouveaux.

## 2.2. Conception du corpus

La conception du corpus complète inclus un certain nombre d'itérations de processus de sélection, dont la sortie était le meilleur candidat en termes de phonétique riche et balance à l'état actuel des unités non couvertes et de leurs distributions. Le processus de sélection pourrait être décrit comme suit. Un énorme sous-ensemble de données brutes y compris le plus rare/ la plus fréquente découvert unité [1] a été extrait pour obtenir un ensemble de phrases candidates. Grâce à la simplicité et l'efficacité, l'algorithme « greedy » a été adopté à la recherche de la meilleure phrase de candidat (avec le poids le plus élevé) dans ce sous-ensemble. Le poids d'une phrase était la proportion de son nombre d'unité distincte découvert et son nombre d'unité distincte totale. Après chaque sélection, l'ensemble de l'unité découvert a été mis à jour, et le processus de sélection a été répété jusqu'à une contrainte (par exemple, la couverture, condition) était satisfait.

Pour examiner la performance du processus de conception proposé, nous avons réalisé la conception qui a examiné di-tonophones comme unités acoustique avec une contrainte de la taille cible de corpus. La sortie est un nouveau corpus de texte, «SAME», avec une taille similaire (ie 24 164 de nombre de phoneme) que l'ancien - VNSP. La taille limitée du corpus cible était petite et le nombre de di-tonophones apparaissant une fois dans les données brutes était considérable. Si les phrases contenant l'unité acoustique le plus rare ont été examinées en premier, ces phrases incluant ces di-tonophones de seule occurrence auraient été les candidats uniques et donc ont été choisis pour le corpus cible. Par conséquent, les phrases contenant *l'unité de la parole la plus fréquente* ont été choisies comme candidates pour le calcul de poids pour maximiser la couverture du corpus cible. Les résultats montrent que, avec un nombre de syllabes semblables, la couverture du nouveau corpus «SAME» était beaucoup plus élevé que l'ancien. Il n'y avait pas ou petite différence du phonème ou de tonophone couvertures, de grandes distances (environ 17-22%) des autres couvertures d'unité entre ces deux corpus. La couverture di-tonophone du nouveau corpus atteint 52,4%, tandis que celle de l'ancien était seulement de 29,6%. Le corpus VSYL avec une couverture complète de la syllabe a également été conçu pour améliorer la qualité de la synthèse de la parole employant la sélection de l'unité non-uniforme.

Le corpus entrainé cible pour notre système de TTS, le corpus VDTS (Vietnamese DiTonophone Speech), a été conçu avec une couverture complète de di-tonophones car il est nécessaire d'enregistrer toutes les transitions entre deux tonophones. De toute évidence, le corpus VDTS avait une couverture de 100% des phonèmes, tonophones et di-phones. Ses couvertures de initiaux/rimes et de syllabes sont 95,1% et 70,2% respectivement.

## 2.3. Enregistrement et prétraitement de la parole corpus

Un total de 5,338 phrases contenant le corpus VDTS et le corpus VNSP, et quelques autres phrases (appelé VDTO – Vietnamese Di- Tonophone and Others) pour la phase d'évaluation ont été enregistrées par une locutrice non-professionnel native de Hanoi,

âgé de 31 ans (nommé Thu- Trang). Les enregistrements ont été réalisés dans un studio bien équipé comprenant une cabine de chant insonorisées et une station de contrôle au Laboratoire LIMSI-CNRS, Orsay, France. Il y avait 27 sessions d'enregistrement d'une heure pour produire près de 8 heures de parole. La qualité de la parole a été contrôlée pendant les sessions par un superviseur. Après plusieurs séances d'enregistrement, des fichiers audio ont été vérifiés pour assurer la qualité globale et de réduire les erreurs pour les prochaines sessions.

Énoncés dans le corpus de la parole ont été renommés par des codes de phrases et prétraitées pour notre système de TTS. Ils ont été segmentées automatiquement et la force-alignés pour construire un corpus annoté par l'étiqueteuse de EHMM. Toutefois, il existait des bruits de mal annoté, qui ont fait des transitions discontinues entre les syllabes de la voix synthétique préliminaire. Ces bruits de respiration ont été donc semi-automatiquement corrigés pour un corpus annoté finale.

Le corpus de parole VDTS qui a été utilisé pour l'entrainement de VTED contient 3947 énoncés, environ 6,4 heures (384 minutes). La vitesse parlée de VDTS est environ 9,6 phonèmes/s, soit 3,6 syllabes/s, donc environ 25% de moins que la précédente enregistrée par un radiodiffuseur. En moyenne, la locutrice Thu Trang produit une pause perçu tous les neuf syllabes, environ 16% de plus de pauses que le radiodiffuseur.

## 3. Proposer un modèle de phrasé prosodique

Dans la synthèse de la parole à base HMM, caractéristiques prosodiques tels que la fréquence fondamentale F0 ou la durée peuvent être entrainées à la fois dans le contexte phonémique et tonale. Le problème restant dans l'analyse prosodique est le phrasé prosodique, le processus d'insertion de pauses prosodiques dans un énoncé. Il comprend l'insertion de pause et des niveaux plus bas de groupement des syllabes. Dans un système de TTS basé à HMM, une pause est considérée comme un phonème; donc sa durée peut être modélisée. Cependant, l'apparition de pauses ne peut pas être prédite par HMM. Le phrasé des niveaux plus bas que les mots ne peuvent pas être complètement modélisé avec des paramètres. A notre connaissance, il n'y a pas d'un tel travail sur la langue vietnamienne. À cause de la contrainte avec les tons lexicaux, l'intonation de niveau phrase dans la langue vietnamienne peut être moins importante dans le phrasé prosodique que dans d'autres langues d'intonation (par exemple, anglais, français). Dans cette recherche, nous avons focalisé au phrasé prosodique pour la synthèse de la parole à partir du texte en langue vietnamienne en utilisant seulement des indices de durée pour deux niveaux de phrasé prosodique pour le TTS du vietnamien: (i) l'apparence de pause - l'un des niveaux les plus importants et fréquents, et (ii) allongement finale .

Dans une étude préliminaire, des règles syntaxiques entre des éléments constituants ou dépendants ont été proposés pour prévoir trois niveaux de pause après un processus de raffinement itératif. Certains traitements statistiques de durée de la pause et l'allongement finale aux frontières prévues ont été réalisés pour raffinage des règles dans le petit corpus VNSP (630 phrases avec annotations manuelles). Ces règles

syntaxiques pourraient bien travailler, soit P = 91,2% et F-score = 69,7%, dans l'environnement manuel, mais seulement des règles syntaxiques constituants ont donné un résultat acceptable (c.-P = 84,2%, score F = 39,9%) dans l'environnement automatique. En outre, avec un analyseur syntaxique automatique, de niveaux de pause prévus ne diffèrent pas significativement conduisant à un seul niveau de pause, soit l'apparence de pause.

Une autre approche, qui a finalement été mis en œuvre pour la version finale de VTED, a été proposée d'utiliser des blocs syntaxiques pour prédire l'allongement final et l'apparence de pause. Blocs syntaxiques ont été proposés comme des phrases syntaxiques avec un nombre attaché de syllabes. L'analyse corpus était le corpus VDTO y compris 5,338 énoncés dans environ 7,7 heures. Les fichiers audio dans ce corpus ont été segmentés automatiquement au niveau de phonème par l'étiqueteuse EHMM. Phonèmes ont ensuite été regroupés aux syllabes et pauses perçus dans un niveau différent. Les fichiers texte ont été automatiquement analysés en arbres de syntaxe par le VTParser, l'analyseur syntaxique Vietnamienne adoptée utilisant l'analyse de shift-réduire avec perceptron moyenne.

Un pattern de durée normalisée (Zscore) dans les syllabes de blocs syntaxiques a été trouvé similaire à ceux des groupes d'haleine (contenant syllabes entre deux pauses perçues consécutives): (i) léger raccourcissement à la première syllabe, (ii) le raccourcissement à l'avant-dernière et (ii) fort degré d'allongement à la dernière. L'allongement final existait encore mais avec un moindre degré dans les dernières syllabes de blocs syntaxiques, excluant les dernières syllabes des ancêtres de niveau supérieurs et des groupes de souffle. En conséquence, en utilisant des indices de durée seuls, les deux niveaux de phrasé prosodique ont été identifiés: (i) apparition de pauses en utilisant des blocs syntaxiques avec un maximum de 10 syllabes et (ii) un allongement final à l'aide de blocs syntaxiques avec un maximum de 6 syllabes.

Des améliorations ont été effectuées par des stratégies de regroupement des blocs syntaxiques simples pour l'allongement final. La prédiction de l'apparition de pauses a été améliorée en combinant les blocs syntaxiques avec deux prédicteurs supplémentaires: (i) de lien syntaxique, une relation / distance d'arbre syntaxique entre deux mots grammaticaux; et (ii) POS. Certaines règles ont été mises en place pour prédire l'apparence de pauses. La performance du modèle fondé sur des règles avec les trois prédicteurs était bonne, à savoir P = 84,8% et F-score = 65,8% lors de la phase d'analyse; P = 84,9% et F-score = 67,4% lors de la phase de test.

Un modèle prédictif a été expérimenté dans une "10-fold cross validation" avec ces trois facteurs prédictifs utilisant J48 (ie l'implémentation Java de l'algorithme C4.5, une approche d'arbre de décision pour le problème de classification). Le "bloc syntaxique" était le facteur prédictif le plus important, car le modèle avec seulement ce prédicteur avait les meilleurs valeurs de Precision (83,4%) et Recall (71,1%), comparativement aux modèles avec seulement POS (F-score = 43,6%) ou un lien syntaxique (F-score = 52,6%) seul. Le "lien syntaxique" prédicteur a aidé le modèle à améliorer la Recall (6%

amélioration), tandis que "POS" a donné des informations efficaces pour augmenter la précision (4% amélioration). Le modèle de composition complet, incluant les trois prédicteurs, avait les meilleurs résultats avec Precision = 89,0%, Recall = 74,6%, donc F-score = 81,2%. En utilisant un ensemble de tests séparés (VDTO-Testing), les performances des deux modèles étaient légèrement différentes. La précision du modèle prédictif était un peu plus élevée que celui à base de règles (soit près de 3%). Cependant, le Recall du modèle prédictif considérablement amélioré, était d'environ 14% plus élevé que celui fondé sur des règles. Ce modèle pourrait être construit automatiquement pour tout locuteur ou tout dialecte par des techniques d'apprentissage machine, et a donc été choisi pour la version finale de VTED.

De nouvelles fonctionnalités d'apprentissage prosodiques ont été proposées pour les systèmes TTS à base HMM vietnamiens. Les évaluations ont montré que la perception de la voix formée avec les nouvelles caractéristiques prosodiques proposées était préférable d'environ 64% - 70% à celle formée sans ces nouvelles fonctionnalités.

## 4. Concevoir et construire VTED

Puisque la motivation initiale était de construire un système TTS de haute qualité pour les aveugles, il doit être complet. Beaucoup d'efforts ont été fournis sur la conception et la construction de VTED, un système TTS HMM vietnamien complet. L'architecture de VTED est composée de trois parties: (i) traitement du langage naturel (NLP), (ii) apprentissage, et (iii) Synthèse. D'après le texte d'entrée, la partie PNL extrait les caractéristiques contextuelles à la fois pour la phase d'apprentissage et la phase de synthèse. Dans la phase d'apprentissage, ces facteurs ont été alignés avec les étiquettes des unités de parole et formés à partir de paramètres de la parole (spectraux et d'excitation) pour construire un modèle HMM dépendant du contexte. Dans la phase de synthèse, conformément à une séquence d'étiquettes de ces facteurs, les caractéristiques contextuelles ont été utilisées pour produire une séquence de paramètres de parole. Enfin, un discours synthétique a été obtenu en utilisant ces paramètres de parole et un vocodeur.

Des fonctionnalités contextuelles pour les Vietnamiens ont été choisies au niveau du tonophone, de la syllabe, du mot, de la phrase et de l'énonciation, héritant d'autres travaux avec adaptation pour les Vietnamiens. Dans une structure stable composée de quatre éléments, bien que le noyau soit obligatoire, l'apparition d'autres éléments est également fréquente. Pour capturer le contexte de tonophones entre les éléments à l'intérieur d'une syllabe ainsi que les transitions aux syllabes précédent / suivant, deux tonophones précédents et deux tonophones suivants ont été choisis pour le contexte phonémique du tonophone actuel. Environ 84% des mots vietnamiens sont des mots composés, dont 70% ont deux syllabes, seulement 1% ont plus de quatre syllabes. Par conséquent, seules une syllabe précédente et une syllabe suivante, ou des niveaux supérieurs ont été considérés comme des caractéristiques de leurs contextes. Il y avait des facteurs locatifs du tonophone, de la syllabe, du mot ou de la phrase actuels dans la syllabe, le mot, la phrase ou l'énonciation actuels. Les numéros de deux niveaux

inférieurs dans le modèle 3-gramme ont également été considérés, par exemple au niveau de la parole: le nombre de tonophones dans le mot précédent / courant / suivant, et le nombre de syllabes dans le mot précédent / courant / suivant. Nous avions des caractéristiques sur les fonctions de phonèmes dans les structures de syllabes (« onset » ou coda), sur les ponctuations et sur certaines caractéristiques prosodiques (par exemple POS du mot précédent / courant / suivant, « break index » des phonèmes, type de position des syllabes). Le ton lexical de la syllabe précédente / actuelle / suivante était une caractéristique importante en vietnamien.

Après avoir étudié attentivement les différentes plateformes, Mary TTS a été choisi pour la construction de VTED en raison de sa facilité d'utilisation, de son évolutivité, et de son vocodeur de haute qualité. C'est une plate-forme de synthèse Text-to-Speech open-source, multilingue et avec une grande communauté de développement dans GitHub. Cette plate-forme facilite la construction d'un système TTS pour une nouvelle langue avec un processus bien conçu.

Beaucoup de travaux ont été réalisés sur la construction d'une partie séparée de traitement du langage naturel pour VTED. Bien que certains modules aient été adoptés à partir d'autres travaux, ils ont nécessité non seulement des efforts, mais aussi du temps pour l'intégration (POS tagger, syntaxe analyseur), l'adaptation et l'extension (Word segmentation), ou même le re-développement (normalisation de texte). Beaucoup de temps a également été consacré pour les dictionnaires de construction, par exemple, le dictionnaire syllabique transcrit PRO-SYLDIC et MEA-SYLDIC, le dictionnaire de mots d'emprunts, le dictionnaire d'abréviations. Les deux modules initiaux étaient la Conversion G2P et la modélisation de prosodie. D'autre part, la plupart des modules Mary TTS existants ont dû être adaptés et modifiés pour les langues tonales, en particulier pour un grand nombre de tonophones.

## 5. Évaluation du système TTS

Plusieurs évaluations perceptives dont des tests MOS, d'intelligibilité, de préférence par paires et d'intelligibilité de tonalité ont été menées pour évaluer la qualité de VTED et le modèle proposé. Grâce à la conception spéciale de certains tests, un outil de test, **VEVA,** a été développé. Ce fut un outil portatif, qui pourrait être déployé dans tout système d'exploitation puisque tous les aspects spécifiques à la plateforme ont été considérés lors de la conception et de la mise en œuvre. Cet outil de test a été déployé et utilisé à la fois sous Windows et Mac OS X pour les tests ci-dessus. L'interface utilisateur graphique (GUI) des tests dans VEVA a été conçue avec un mode plein écran, ce qui a empêché les sujets de se laisser distraire par d'autres objets (par exemple, des icônes, des applications et des barres) sur l'écran.

Tous les tests de perception ont eu lieu au Vietnam. Certains traitements statistiques ont été effectués sur les résultats des tests pour confirmer leur fiabilité. Les résultats ont montré que la perception de la dernière version de VTED approchait de celle de la parole naturelle, en termes d'intelligibilité et d'intelligibilité du ton lexical. Le naturel de la

dernière version a considérablement progressé au cours de la première mise en œuvre, et a eu un léger écart par rapport à la parole naturelle.

Le test initial MOS a été mené sur la première version de VTED (appris avec l'ancien corpus VNSP) et sur le système précédent, HoaSung, utilisant la sélection de l'unité non uniforme avec le même corpus d'entraînement, et de la parole naturelle comme référence. Les résultats ont montré que cette version initiale de VTED était plutôt bonne, 0,81 (sur une échelle de 5 points MOS) plus élevé que HoaSung. Toutefois, le score du premier VTED était encore de 1,2 points de moins que la parole naturelle. Dans le test final MOS, les résultats de l'expérience ont montré que la qualité de la version finale de VTED (appris avec le nouveau corpus VDTS et le modèle de phrasé prosodique proposé avec des blocs syntaxiques) avait progressé d'environ 1,0 (sur une échelle MOS de 5 points) par rapport à sa première version. Le modèle proposé de phrasé prosodique a été soutenu par VTParser, un analyseur syntaxique automatique pour les Vietnamiens. L'écart entre le discours synthétique de VTED et de la parole naturelle a aussi beaucoup diminué. Bien que les résultats absolus de la première version de VTED dans les tests MOS initiaux et finaux furent différents (soit 3,6 et 2,9 sur une échelle MOS de 5 points), les écarts relatifs entre les voix dans les deux tests sont restés comparables. Le score de première version de VTED était d'environ 1,2 points lors du test initial, et d'environ 1,5 points lors du test final ; ces scores sont inférieurs à ceux obtenus par la parole naturelle. Le score de la voix naturelle dans le test final était d'environ 0,3 points de moins que celui du test initial. Il s'est avéré que les participants au test final ont donné des les taux plus "strictes" et "sensibles" que dans le premier.

Les tests perceptifs ont montré qu'avec de nouveaux modèles de phrasé prosodiques, le discours synthétique de VTED a été préféré d'environ 64% (analyse automatique et blocs syntaxiques) ou d'environ 70% (analyse manuelle et règles syntaxiques) sur la version précédente. Les deux évaluations subjectives et objectives ont été effectuées pour confirmer que les caractéristiques Tobi n'améliore pas la qualité de systèmes TTS vietnamiens en général, et la dégrade même dans certains cas. En conséquence, les caractéristiques de TOBI ont été retirées de la deuxième version de VTED.

Le test d'intelligibilité a été conçu avec une matrice en carré Latin 3x3 pour trois voix: la première version de VTED, la version finale de VTED, et un discours naturel. Au niveau des syllabes, le taux d'erreur de la première version de VTED était d'environ 14,3%, soit environ 12,0% supérieur à celui du discours naturel. Cette première version a divergé d'environ 5.8 à 7.5% par rapport la parole naturelle de niveaux inférieurs : le ton et le phonème. L'écart entre la dernière version de VTED et le langage naturel n'est que de 0,4% - 1,4%. Ce résultat a montré que la dernière version de VTED a considérablement avancé en termes d'intelligibilité, qui s'approche de celle de la parole naturelle.

Dans l'intelligibilité du ton lexical, des groupes de phrases significatives avec les mêmes syllabes et le même ordre de syllabe, divergentes par un seul ton, ont été préparés. Les sujets ont été invités à choisir la syllabe la plus probable qu'ils avaient entendue parmi un groupe de syllabes portant différents tons dans un énoncé. Dans le premier test, environ 23% en moyenne et - selon le type de tonalité – une différence de 0% à

37% par rapport à la parole naturelle a été perçue. Les écarts entre le premier VTED et la parole naturelle dans les deux tests initiaux et finaux sont restés similaires, soit environ 22% dans le test initial, et 21% dans le test final. Nous croyions que ce genre d'évaluation perceptive pourrait être considéré comme une évaluation «absolue» car les sujets ne doivent pas donner de «score», mais seulement choisir la syllabe la plus probable qu'ils avaient entendue. Dans le test final, la dernière version de VTED a reçu des taux de bonnes réponses élevés, de 96% à 100%, pour tous les tons sauf pour le ton descendant 2 - seulement 76% perçus correctement. Le taux de bonnes réponses global de la version finale de VTED dans l'identification des tons en contexte était seulement 2,6% inférieur à celui de la parole naturelle. Il s'est avéré que l'intelligibilité de du ton lexical de la dernière version de VTED a également approché celui de la parole naturelle.

# 6. Perspectives

Cette recherche a abouti à plusieurs réalisations vers la construction d'un système de TTS de haute qualité pour les Vietnamiens. Toutefois, la poursuite des travaux peut être faite afin d'améliorer la qualité du système, ainsi que d'élargir la recherche à une gamme d'applications. Certaines grandes perspectives de ce travail sont présentées ici.

## 6.1. Amélioration de la qualité de la voix synthétique

La qualité de la voix synthétique de la dernière version de VTED était bonne, et a même atteint le discours naturel en termes d'intelligibilité générale et du ton. Toutefois, elle était encore nettement distinguée de la parole naturelle dans le test MOS (environ 0,5 point de moins que la parole naturelle sur une échelle de 5 points). Dans l'intelligibilité de la voix, bien que la plupart des tons ont reçu des taux de bonne réponse élevés (de 96% à 100%), identiques à ceux du langage naturel, le ton descendant 2 a été faussement identifié d'environ 24%. Avec un petit corpus, le taux d'erreur de la voix de synthèse est de 23% en moyenne et variait de 0% à 37% selon le type de tonalité. Ce résultat montre que les tons lexicaux ne sont pas bien modélisés pour des cas spécifiques, même avec un bon corpus.

La justesse de ton de la voix de synthèse peut être améliorée par l'étude des structures de l'arbre de décision dans le processus de regroupement de contexte fondé sur un arbre de la formation HMM. D'autre part, un corpus d'enregistrement de signaux issus d'un électroglottographe (ci-après EGG) peut aider à améliorer l'intelligibilité de la voix et de l'intonation de la voix synthétique. Avec les signaux issus de l'EGG, la F0 peut être calculée de manière plus précise et plus fiable qu'avec des algorithmes traditionnels (par exemple l'algorithme ESPS utilisant l'outil « snack » dans ce travail). Puisque la F0 est un paramètre essentiel de la parole (paramètre d'excitation) dans la formation basée-HMM, nous croyons que plus la F0 est extraite avec précision, plus la qualité de la parole générée s'améliore.

La précision des modèles acoustiques peut être améliorée par une nouvelle technique d'apprentissage de premier plan, tels les réseaux neuronaux profonds (DNN). Par exemple, l'utilisation de DNN pour l'arbre de décision peut répondre à certaines limites de l'approche classique tel que l'inefficacité dans l'expression de dépendances de contexte complexe, la fragmentation des données de formation, et l'ignorance complète des caractéristiques d'entrée linguistiques qui ne figuraient pas dans les arbres de décision.

### 6.2. TTS pour d'autres dialectes vietnamiens

Bien que le vietnamien de Hanoi moderne soit considéré comme la norme vietnamienne, il y a toujours un besoin d'étendre VTED afin qu'il soit capable de synthétiser d'autres dialectes vietnamiens populaires pour les résidents locaux. Trois autres dialectes principaux sont Hue, Nghe An (Centre du Vietnam), et le sud du Vietnam.

Les deux tâches principales pour développer un système de TTS pour un dialecte vietnamien comprend: (i) l'étude de la phonétique et de la phonologie de ce dialecte, (ii) la conception et l'enregistrement d'un nouveau corpus pour celui-ci. Ces tâches peuvent réutiliser les résultats des travaux en cours et faire une adaptation pour le dialecte cible. Les nouveaux corpus peuvent ensuite être automatiquement segmentés, étiquetés et formés pour une nouvelle voix de ce dialecte.

### 6.3. Synthèse de la parole expressive

La synthèse de la parole expressive a été considérée pour accélérer la qualité des systèmes de synthèse vocale à une nouvelle vision, ce qui aidera les utilisateurs à accepter les signaux générés par le TTS grâce à des discours moins impersonnel. Par exemple, un système TTS peut générer un son de voix heureuse ou morose, amicale ou empathique, autoritaire ou incertaine. Il existe plusieurs directions de recherche sur ce sujet, tels que (i) la caractérisation du haut-parleur et le personnalisation de la voix: les modèles qui peuvent être adaptés à un locuteur, ce qui permet de tenir compte de leur humeur, de leur personnalité ou de leurs origines, (ii) modélisation de la durée et de la prosodie: le contrôle de la prosodie (par exemple, durée phonème, le contrôle de la mélodie) pour obtenir une voix adaptée au contexte de l'interaction.

Avec la synthèse de la parole expressive, nous pouvons générer automatiquement des livres audio de haute qualité dans lesquels différents styles de lecture peuvent être adaptés à différents personnages. Différentes voix peuvent être utilisés pour générer des dialogues pour les différents personnages. La synthèse vocale émotionnelle peut capturer une multitude d'expressions émotionnelles (par exemple colère, peur, joie et tristesse) de chaque personnage dans les dialogues. De même, les sous-titres dans les films peuvent également être automatiquement convertis en parole expressive avec peu d'effort.

### 6.4. Voice Reader

Le lecteur Sao Mai vietnamien est un système de synthèse par concaténation avec un corpus dont les syllabes ont été enregistrées séparément. En dépit de la faible qualité et

d'autres inconvénients, le lecteur vietnamien Sao Mai est actuellement considéré comme le logiciel le plus commun de soutien aux aveugles vietnamien à utiliser des ordinateurs personnels. Bien qu'il existe plusieurs autres lecteurs vietnamiens ciblant les utilisateurs aveugles, ils ne sont pas utilisés dans la vie réelle en raison de leurs limitations d'utilisabilité.

À partir de la conception et de la mise en œuvre de VTED, il est nécessaire de poursuivre les travaux pour créer un tel lecteur. Tout d'abord, quelques optimisations dans la conception du corpus ainsi que des fonctionnalités contextuelle devraient être menées. Deuxièmement, les interactions entre le système TTS et son environnement doivent être améliorées. Enfin, les fonctions nécessaires d'un lecteur doivent être étudiées et mises au point sur la base des exigences de l'utilisateur.

## 6.5. Machine de lecture

Le système TTS peut également être appliqué aux systèmes embarqués spécialisées pour la lecture de la parole à partir d'autres formats, tels que le texte ou l'image. Il peut même être combiné avec un scanner et une application OCR pour fournir aux utilisateurs une entrée de données à partir des enregistrements de données de papier.

Pour la langue vietnamienne, cette machine sera très utile non seulement pour les aveugles et les personnes ayant une basse vision, mais aussi pour tout utilisateur qui veut écouter des e-journaux, journal, ou même des messages sans lire. Basé sur le système de TTS actuel vietnamiens, un système de TTS léger doit être conçu et développé sur les systèmes embarqués. "Flite + hts_engine" peut être utilisé comme une bonne référence pour un tel système.