Script optimization for TTS voice corpus design in audio-book generation

Meysam Shamsi

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : Informatique

Par

Meysam SHAMSI

«Script optimization for TTS voice corpus design in audio-book generation»

Thèse présentée et soutenue à « Université de Mans », le «16 octobre 2020 »
Unité de recherche : UMR CNRS 6074 - IRISA
Thèse N° :

Rapporteurs avant soutenance :

Frédéric BÉCHET	Professeur HDR, Aix Marseille Université
Slim OUNI	Maître de Conférences HDR, Université de Lorraine

Composition du Jury :

Président :	Sylvain MEIGNIER	Professeur HDR, Université du Mans
Examinateurs :	Frédéric BÉCHET	Professeur HDR, Aix Marseille Université
	Elisabeth DELAIS-ROUSSARIE	DR CNRS HDR, Université de Nantes
	Sylvain MEIGNIER	Professeur HDR, Université du Mans
	Slim OUNI	Maître de Conférences HDR, Université de Lorraine
Dir. de thèse :	Damien LOLIVE	Maître de Conférences HDR, Université de Rennes1

Invités :

Nelly BARBOT	Maître de Conférences, Université de Rennes 1 - IRISA
Jonathan CHEVELU	Maître de Conférences, Université de Rennes 1 - IRISA

ACKNOWLEDGEMENT

This thesis has been realized under the ANR (French National Research Agency) project SynPaFlex ANR-15-CE23-0015 and also funded by the Région Bretagne and the Conseil Départmental des Côtes d'armor.

I would like to thank Damien Lolive my thesis director for his long term guideline and priceless comments on my works and ideas, Jonathan Chevelue for his great ideas and technical helps, Nelly Barbot for her appreciable help in writing, organization and formulating the ideas.

My gratitude goes to Frédéric Béchet and Slim Ouni who reviewed my thesis, and to Elisabeth Delais-Roussarie and Sylvain Meignier for taking part in my committee.

A very very special thanks to all of the members in Expression group, for their help in my research and fruitful discussions. I have also been lucky enough to have some amazing friends throughout this PhD journey. A special note of thanks to Cédric Fayet, Aghilas Sini, Antoine Perquin, Somayeh Jafaritazehjani, Valentin Durand-De-Gevigney.

The last years of my PhD would not have been possible without the support of my wife, Shamim. Finally, the most important people responsible for any success I've had are my mother and father. They have always put my needs far above their own, and have always been supportive of whatever I decision I've made. I sincerely owe a lot to them.

TABLE OF CONTENTS

RÉSUMÉ EN FRANÇAIS

Ce résumé est une version condensée en français de l'ensemble des considérations, hypothèses et expérimentations, agrémentées de leurs résultats, présentées en langue anglaise dans cette thèse. Un soin particulier a été apporté au respect de l'ordre de présentation des idées développées dans le manuscript en anglais et dans le présent résumé, de sorte que chaque chapitre du premier correspond à une section du dernier.

Dans cette thèse, nous proposons de réduire le coût de génération d'un livre audio en synthétisant une partie du livre à l'aide d'un système de synthèse de parole (système TTS, abbréviation de Text-to-Speech). Afin d'avoir un style cohérent et d'offrir la meilleure qualité possible, il est proposé qu'une partie du livre soit lue et enregistrée par un locuteur professionnel et utilisée comme corpus vocal pour synthétiser le reste du livre. Le livre audio final serait donc une combinaison de signaux vocaux naturels et synthétiques. Le problème d'optimisation associé consiste à sélectionner le script d'enregistrement le plus court possible pour générer le livre audio avec une qualité aussi élevée que possible.

Afin de pouvoir procéder à des évaluations de différentes méthodes de sélection d'un script et de caractéristiques associées telles que sa longueur, le processus d'enregistrement est "simulé" par l'utilisation d'un livre entièrement enregistré au préalable. En effet, le coût élevé associé à un enregistrement limite le nombre d'évaluations en situation réelle. Outre ce coût, celui des tests de perception est un autre défi auquel nous sommes confrontés dans cette thèse. Les mesures objectives pour l'évaluation de la solution du sous-ensemble ne peuvent être qu'une approximation de la qualité finale. Toute évaluation objective doit être confirmée par des tests de perception.

Contexte

Le chapitre 1 rappelle le contexte scientifique du sujet de thèse. Cela commence par la définition de certains termes dans ce domaine et se poursuit par une brève introduction des systèmes de synthèse vocale et de l'évaluation de la qualité synthétique

de la parole.

Les différences entre le texte et la voix qui doivent être prises en compte dans les systèmes TTS sont précisées. Les termes expressivité et qualité de la parole sont également définis. Trois principaux systèmes TTS sont présentés brièvement ; le système statistique paramétrique, le système basé sur la sélection d'unités et l'approche hybride qui combine des éléments des deux précédents pour associer leurs avantages respectifs. La synthèse de parole expressive est alors introduite, avant que ne soit abordée l'évaluation de la qualité vocale de signaux synthétiques, de manières objective et subjective.

Conception de corpus vocal

Le chapitre 2 présente le problème de la conception d'un corpus vocal. L'objectif de la thèse est formalisé comme un problème de partitionnement d'un livre en deux parties (a priori de taille inégale) afin de le vocaliser de la façon suivante : l'une des parties correspond alors à des signaux de parole naturelle constituant une voix à partir de laquelle la seconde partie est vocalisée de manière synthétique avec la meilleure qualité possible.

Basée sur la littérature, la conception du corpus est étudiée dans les sections 2.2 et 2.3 portant sur la réduction du corpus vocal et la sélection d'un script d'enregistrement. En effet, un corpus vocal existant doit parfois être réduit afin d'augmenter son adéquation, en tant que voix alimentant un système TTS, au contexte de l'application. D'autre part, la conception d'une voix peut être traitée en sélectionnant un script textuel dont la lecture sera enregistrée. Dans cette approche, le script doit être aussi court que possible afin de minimiser le coût humain d'un processus d'enregistrement et d'étiquetage de haute qualité. Ainsi, cette dernière approche est proche de notre problème d'origine. Les travaux précédents sur la sélection de texte en fonction de la couverture des unités ou de leur distribution sont présentés.

De la réduction du corpus à la sélection du script

Afin d'adapter la sélection de script à notre tâche TTS, l'étude préalable du processus de réduction du corpus vocal et de son impact pourrait être utile pour identifier

les caractéristiques requises à l'optimalité d'un script d'enregistrement. Le chapitre 3 étudie le problème de réduction du corpus vocal TTS.

Tester tous les sous-ensembles possibles afin d'en déterminer le meilleur est impossible compte-tenu de la combinatoire associée. Cette dernière est donc le principal défi de ce problème de sélection d'un sous-ensemble optimal. En modifiant un glouton cracheur standard, le temps de calcul par l'algorithme est réduit à un temps raisonnable. Cependant, l'heuristique adoptée nuit à la qualité du résultat, le rapprochant d'une sélection aléatoire.

Dans un premier temps, certaines mesures objectives comme PESQ, DTW entre les signaux de parole synthétiques et naturels ainsi que le coût global associé à la sélection des unités sont étudiées. Un test d'écoute montre que le coût global TTS a une corrélation plus forte avec la qualité perceptuelle. De plus, en considérant différentes mesures pour procéder au classement des candidats à chaque itération de l'algorithme glouton, aucune supériorité significative des mesures linguistiques telles que l'entropie et la KLD sur le coût global TTS n'est observée.

À l'aide d'un test MUSHRA, on constate qu'au delà d'une certaine taille de voix (1 heure de notre livre audio), l'augmentation de cette voix n'entraîne pas de gain significatif quant à la qualité des signaux synthétiques évalués perceptuellement. De plus, aucune différence entre la réduction aléatoire et la réduction gloutonne proposée n'a été observée. Afin d'évaluer la performance de ces deux approches, un autre test de préférence AB est effectué, confirmant l'absence de préférence des auditeurs.

Pour résumer, nous n'avons pas trouvé d'algorithme, qui en un temps de calcul raisonnable fonctionne mieux que le hasard, nous obligeant à abandonner l'analyse a posteriori des caractéristiques d'une voix réduite et optimale initialement envisagée. Malgré cela, le coût global TTS s'est révélé être une bonne mesure objective pour estimer la qualité perceptuelle d'un signal synthétique.

Approche basée sur le plongement de phonèmes

Le chapitre 4 a pour objet la construction d'une voix au moyen de la sélection d'un script. L'objectif est d'extraire un sous-ensemble du livre ciblé à l'aide d'informations linguistiques. Afin de représenter de manière continue ces informations et faciliter la sélection d'un sous-ensemble d'énoncés, différentes approches de plongement de phonèmes sont comparées.

Un réseau neuronal à convolution profonde (DCNN) est utilisé pour projeter des informations linguistiques dans un espace d'intégration. La représentation ainsi obtenue corpus est ensuite exploitée par un processus de sélection pour extraire un sous-ensemble d'énoncés offrant une bonne variété linguistique tout en tendant à limiter la répétition d'unités linguistiques. Nous présentons deux processus de sélection: une approche de clustering (K-Means) basée sur la distance d'énonciation et une autre méthode qui tend à atteindre une distribution cible d'événements linguistiques (basée sur KLD).

Les expériences montrent qu'un auto-encodeur CNN peut être utilisé avec succès pour extraire des informations linguistiques. Le clustering K-Means et les méthodes KLD fonctionnent correctement en utilisant des représentations intégrées qui obtiennent de meilleurs résultats qu'une approche aléatoire, ou même que des méthodes classiques telles que l'algorithme glouton pour des couvertures d'ensembles dans des espaces de recherche discrets. Nous avons également comparé les trois approches d'intégration CNN, LSTM et *Doc2Vec*, et CNN s'avère mieux adaptée dans ce contexte particulier de la conception de corpus vocal. L'évaluation subjective a confirmé ce résultat montrant plus généralement une préférence pour les approches à base de plongement proposées.

Modèle acoustique et conception du corpus vocal

Le plongement d'informations linguistiques peut être utilisé à différentes tâches de traitement de langage naturel. Le chapitre 5 considère ainsi des modèles acoustiques basés sur un plongement des informations linguistiques pour les associer aux informations acoustiques comme solution pour la conception de corpus vocaux TTS. Dans un second temps, la relation entre la conception des corpus vocaux et la synthèse hybride y est étudiée.

Tout d'abord, trois architectures différentes pour une modélisation conjointe des informations linguistiques et acoustiques sont proposées pour la conception de corpus vocaux. Leurs signaux synthétiques ont été comparés à ceux obtenus à partir de la voix issue du meilleur modèle de plongement de phonèmes, *CNN-KMeans*, étudié au chapitre 4. L'évaluation objective effectuée à l'aide du coût global TTS n'a montré aucune supériorité du modèle acoustique par rapport au modèle *CNN-KMeans* pour la conception du corpus.

La relation entre la conception du corpus vocal et le système TTS basé sur la sélection d'unités a été étudiée. Il s'avère, dans le cadre d'un système TTS hybride, que l'utilisation d'un modèle de plongement acoustique pouvait surpasser le modèle de plongement linguistique CNN proposé en tant que fonction de coût cible, bien que la voix ait été conçue par l'approche *CNN-KMeans* et que le corpus d'apprentissage du modèle acoustique soit plus petit que celui du modèle linguistique. On observe par ailleurs que l'intégration du plongement linguistique dans le système TTS hybride produit des signaux de meilleure qualité que le système TTS "expert" (sélection d'unités sans hybridation). A contrario, la prise en compte d'informations acoustiques n'améliore pas le processus de sélection des scripts.

Les phrases les plus courtes

Certaines études soulignent que les algorithmes de réduction ont tendance à sélectionner des énoncés plus courts, de manière plus ou moins importante selon les corpus initiaux. La méthode *CNN-KMeans* non supervisée de notre expérience précédente sélectionne également des phrases de longueur plus courte que la longueur moyenne. Sur la base de ces observations, nous proposons une approche consistant à simplement sélectionner les phrases les plus courtes du livre. Le chapitre 6 compare cette méthode de conception de corpus vocaux TTS avec trois autres méthodes précédemment étudiées. Pour cela, deux types de systèmes TTS par concaténation et deux livres audio en langue française avec des longueurs et des styles d'énoncés différents sont utilisés.

Les résultats expérimentaux montrent qu'une méthode simple comme la sélection d'énoncés courts pourrait bien fonctionner pour la conception de voix pour la synthèse de parole pour la génération de livres audio lorsque le corpus vocal est une partie du livre. Pour les deux livres audio, que ce soit pour le système TTS expert ou hybride, cette méthode fonctionne mieux que l'approche *CNN-KMeans*.

Pour une voix de taille suffisamment importante, on observe que la voix résultant d'une approche classique gloutonne pour couvrir les di-phonèmes ne produit de meilleurs signaux qu'une voix construite aléatoirement. En comparant le coût global TTS, le taux de couverture des di-phonèmes et tri-phonèmes ainsi que la divergence de Kullback-Liebler entre les distributions linguistiques de la voix et du livre complet, les stratégies basées sur la couverture de di-phonèmes ou sur la mesure KLD ne con-

duisent pas à de meilleures voix. Ces mesures ne sont donc nécessairement une bonne métrique de conception de voix pour la TTS lorsque la taille de celle-ci est suffisamment grande.

Évaluation des signaux mixtes synthétiques et naturels

Le chapitre 7 étudie l'impact de la configuration de signaux de parole naturelle et synthétique dans le livre audio final. Deux facteurs sur la perception des auditeurs sont considérés. Nous comparons des énoncés entièrement vocalisés de manière synthétique avec des énoncés combinant, à parts égales, des signaux naturels et synthétiques agencés dans différents ordres, avec différents niveaux de qualité synthétique.

Un test de perception a montré que les énoncés "mixtes" sont préférés par les auditeurs aux énoncés vocalisés de manière entièrement synthétique. Cela a été observé pour les niveaux de qualité synthétique testés (déterminés par la taille de la voix). Ce résultat confirme la première hypothèse de la thèse selon laquelle voix alimentant le système TTS devrait faire partie du livre audio final. Par conséquent, enregistrer une partie du livre audio et l'utiliser pour synthétiser le reste du même livre aiderait à avoir une qualité globale supérieure dans le livre audio final au lieu de tout synthétiser.

Quelle que soit la proportion de signaux synthétiques dans un livre audio mixte, le changement de type de signal peut influer sur la perception des auditeurs. Par conséquent, l'impact du nombre de transitions dans les vocalisations mixtes, moitié synthétiques et moitié naturelles, a été étudié. Les scores MOS et une comparaison directe dans un test AB ne montrent pas que le nombre de transitions pourrait changer la perception et la préférence des auditeurs. Le test AB devait à l'origine étudier l'effet de la longueur de la partie synthétique continue dans le signal mixte sur la perception des auditeurs. À proportion égale entre signaux naturels et synthétiques, les évaluateurs n'avaient aucune préférence entre une vocalisation avec signal synthétique long, et celles utilisant deux signaux synthétiques courts.

Nos analyses des résultats montrent que la perception des auditeurs est influencée par la nature des signaux au début et à la fin des vocalisations mixtes (4 groupes de souffle) : les signaux mixtes commençant par une partie naturelle et se terminant par une partie synthétique sont préférés. Cependant une évaluation de l'impact de la position des signaux de différentes natures reste nécessaire sur des énoncés de longueur plus importante (à l'échelle du paragraphe, chapitre, etc.). Malheureusement,

en raison de la durée et pénibilité des tests perceptuels, l'évaluation de signaux plus longs limite le nombre d'échantillons testés par auditeur et nécessite plus d'auditeurs.

Conclusion

L'objectif principal de cette thèse est la conception de voix pour la synthèse de parole TTS par sélection d'unités dans la tâche de génération de livres audio. Dans ce cadre, le livre audio final est un mélange de signaux de parole synthétiques et naturels. La partie synthétique est produite par un système TTS alimenté par les signaux de parole naturels correspondant à la lecture enregistrée de l'autre partie du livre. La sélection de la partie à enregistrer comme voix pour la TTS est la principale préoccupation de cette thèse.

Le premier outil nécessaire à la conception de voix TTS est une mesure d'évaluation objective. Il a été montré que le coût global TTS offre une bonne corrélation avec les scores d'évaluation perceptive de la qualité des signaux synthétiques.

Une méthode entièrement non supervisée qui peut prendre en compte les informations contextuelles et plonge les informations linguistiques discrètes dans un espace continu est présentée. Ce modèle de plongement se révèle efficace pour la sélection de scripts dans le problème de conception de voix.

Une analyse des résultats de nos premières expérimentations et de la littérature montre une tendance a posteriori à sélectionner à des énoncés courts pour construire une voix pour la TTS. Une expérience sur deux livres audio, avec des longueurs moyennes d'énoncés différentes, et deux systèmes TTS concaténatifs a confirmé qu'une voix composée des énoncés les plus courts est plus efficace que toutes les méthodes précédentes dans le contexte applicatif de la thèse.

Enfin, en terme de qualité globale, des évaluations perceptuelles ont montré qu'une vocalisation mixant signaux de parole naturels et synthétiques est préférée à une vocalisation entièrement synthétique d'un même énoncé. Cependant, nous n'avons pas constaté d'impact du nombre de transitions entre les signaux synthétiques et naturels sur la préférence des auditeurs.

Perspectives

Cette thèse est centrée sur la conception de voix pour les systèmes TTS à base de sélection d'unités. Les résultats obtenus nécessitent d'être confrontés à ceux qui résulteraient de systèmes de synthèse de parole *end-to-end*.

La pertinence de l'approche de conception de voix à l'aide de sélection d'énoncés courts en fonction du niveau d'expressivité souhaité peut être étudiée. La caractéristique acoustique et linguistique des énoncés courts pourrait être le futur sujet d'étude. Alors que les informations acoustiques jouent un rôle important dans le contexte TTS, une étude approfondie serait nécessaire pour identifier les causes de l'absence d'amélioration de la qualité des signaux synthétiques produits à l'aide d'une voix dérivée des modèles acoustiques étudiés dans cette thèse.

Liste des publications

1. Meysam Shamsi, Nelly Barbot, Damien Lolive, Jonathan Chevelu, « Mixing Synthetic and Recorded Signals for Audio-book Generation», in: 22st International Conference on Speech and Computer (SPECOM), Springer, St. Petersburg, Russia, October 2020.
2. Meysam Shamsi, Jonathan Chevelu, Nelly Barbot, Damien Lolive, «Corpus design for expressive speech: impact of the utterance length», in: International Conference of Speech Prosody, Tokyo, Japan, May 2020, pp. 955–959.
3. Meysam Shamsi, Damien Lolive, Nelly Barbot, Jonathan Chevelu, «Corpus Design using Convolutional Auto-Encoder Embeddings for Audio-Book Synthesis», in: Annual Conference of the International Speech Communication Association (InterSpeech), Graz, Austria, Sept 2019, pp. 1531–1535.
4. Meysam Shamsi, Damien Lolive, Nelly Barbot, Jonathan Chevelu, «Investigating the relation between voice corpus design and hybrid synthesis under reduction constraint», in: 7th International Conference on Statistical Language and Speech Processing (SLSP), vol. 11816, Ljubljana, Slovenia: Springer, Cham, Oct 2019, pp. 162–173.
5. Meysam Shamsi, Damien Lolive, Nelly Barbot, Jonathan Chevelu, «Script Selection using Convolutional Auto-encoder for TTS Speech Corpus», in: 21st In-

ternational Conference on Speech and Computer (SPECOM), Springer, Istanbul, Turkey, Aug 2019, pp. 423–432.

6. Meysam Shamsi, «TTS voice corpus reduction for audio-book generation», in: Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition), Nancy, France, Jun 2020, pp.193-204.
7. Meysam Shamsi, Script Optimization for the Expressive Synthesis of Audiobooks, 5th Doctoral Consortium ISCA-SAC, Graz, Austria, Sept 2019.

INTRODUCTION

Audio books have wide varieties of applications. They could be used for disabled people, learning material and entertainment. The conventional way to generate an audio book is to record a professional speaker in a studio. This process is costly, time-consuming with the difficulty of ensuring consistent voice quality throughout the recording phase. A solution to reduce the cost of audio-book generation is the use of a Text-To-Speech (TTS) synthesis system, which converts a given script to a speech signal. TTS systems are widely used in industry nowadays. Recently, they have made great progress in terms of acoustic quality and intelligibility. Nevertheless, some applications still require improvements for further developments, like audio-book generation. Producing a high quality expressive signal for audio-books still remains a research problem. Moreover synthesizing signals with proper para-linguistic features such as emotion, prosody, style and intonation is a key point for audio books in order to be pleasant for listeners.

The speech quality strongly depends on two main principles of speech synthesis. TTS systems have their advantages and weaknesses according to their types. The type of system (parametric or unit selection) can influence different aspects of synthetic speech quality such as naturalness and expressiveness. Besides TTS system type, the synthetic speech quality is also strongly affected by the quality of the voice corpus. In this thesis, the main concern is voice corpus design for a unit selection based TTS. A random or unbalanced corpus contains lots of phonological unit repetitions and, most importantly, does not guarantee a sufficient variety of units for the speech synthesis process. Moreover, the corpus should be as small as possible in order to minimize the human cost of high quality recording and labeling checking stages. In the case of unit selection and hybrid approaches, a reduced corpus size may also accelerate the synthesis process considering the smaller search space. In that case, removing redundant elements while adding critical ones to the corpus is important. A well-designed corpus combines parsimony and balanced unit coverage in order to gain a satisfactory level of richness with a minimal construction cost.

The goal of synthesizing signal in a general task like reading news and vocal as-

sistance is only transforming the message content. Thus, maximizing covering of linguistic and alphabetical variation with minimum corpus length is the main concern in such case. Synthesizing a speech signal in an expressive task like audio-book generation should take into account the para-linguistic information as well. The pleasantness, prosody, and style of synthetic signal could be as important as the message content for audio-book listeners. Moreover, sometimes these para-linguistic features are not considered during the annotation process and should be extracted according to context.

In this thesis, we propose to reduce the generation cost of an audio book by synthesising a portion of the book. In order to have a coherent style and keep the highest possible quality, the rest of the book should be recorded by a professional speaker and used as the voice corpus for synthesising the first part. So the final audio book would be a mix of natural and synthetic speech signals. The optimization problem is to select the best recording script, as short as possible, to generate the audio book with a quality as high as possible.

The extracted subset solution should be first recorded to be evaluated. To simulate the process, a fully recorded book can be used. The recording cost limits the number of evaluations in real situation. Besides the recording cost, the subset selection methods should be evaluated based on the perceptual quality of final audio-book. The cost of perceptual tests is another challenge that we are faced with this thesis. Any objective measure of subset solution quality can only be an approximation of the final quality and objective evaluations should be confirmed by listening tests.

This thesis is organized in four main parts. In the first part (chapters 1 and 2) the terminology, related works and the main thesis objective are described. Chapter 1 reviews the literature, TTS systems and evaluation method for speech quality. The corpus design problem and the state of the art methods are then detailed in chapter 2. In order to investigate the problem and its context, the script selection for TTS voice corpus is revised as TTS voice corpus reduction in the second part (chapter 3). This chapter discusses a posterior strategy to find the best voice subcorpora which can lead future studies. In the third part (chapter 4 and 5 and 6) voice corpus design is investigated based on linguistic features. A phoneme embedding model is proposed in chapter 4. This embedding model is considered as a linguistic feature extraction method and is followed by a selection method. Afterwards, in chapter 5, the linguistic embedding model is replaced by acoustic models and the usage of the phoneme embedding model in hybrid TTS is discussed. By analysing the results of script selection

methods that cause good synthetic quality, the idea of selecting short utterances for TTS voice corpus design is reviewed in chapter 6. Finally in chapter 7, the idea of generating an audio-book as mixed synthetic and recorded signals is studied.

CHAPTER 1

BACKGROUND

This chapter gives a preview of the thesis domain. It starts with the definition of some terms in this field and it goes on with a brief introduction of speech synthesis systems and the evaluation of synthetic speech quality.

The aim of the definition section is to reduce ambiguity of terms that are commonly used in this document. By concentrating on differences between text and voice, an automatic conversion from Text-to-Speech has to challenge. The expressiveness and speech quality will be defined in following to to introduce the main parameters that impact the speech quality. Second section will consider text-to-speech systems. Three kinds of speech synthesis and the state of the art systems will be mentioned shortly. The requirement and special consideration of expressive speech synthesis will be discussed in section 1.3. The evaluation methods and protocols will be reviewed in section 1.4. The perceptual protocols, objective methods and previous works on evaluation modeling will be described.

1.1 Definition of terms

To clarify terminology and before diving into the thesis context, some terms will be defined in this section. The differences between text and voice, the challenge and definition of expressiveness and speech quality will be considered in the following.

1.1.1 Text versus voice

Speaking, writing, and other communication ways are the embodiment of language with different capacities. The Text-to-Speech (TTS) system transform a given script into acoustic signal. There is some standardized representations of the sounds like International Phonetic Alphabet (IPA), however, a given letter can be uttered differently according to context and production source. In (Campbell, 2007), it is noticed that not

only speech is older (in terms of human evolution) than text and writing, but also it has more information types. In order to distinguish between a writing form of a speech and its acoustic representation, two terms will be used in following: *phoneme* and *phone*. A phone is the acoustic realization of a phoneme (the smallest distinguishable part of script to be pronounced).

According to (Fujisaki, 2004), the information expressed by speech can be categorized in three types, though their boundaries may not always be clear. The linguistic information, which is based on written language and text, contain semantic information. The para-linguistic information which is uttered by speaker to specify, modify, or supplement the linguistic information. And the non-linguistic information which is not generally controlled by the speaker, such as the speaker's gender, age, idiosyncrasy, vocal ability, and etc.

However, even if acoustic descriptions are strongly bonded with linguistic definitions, they have some differences. The para-linguistic components including tone, stress, pitch, and volume and even the extra-linguistic information like nonverbal communication (expressions, gestures, and movements) help speech to convey more information. Speech is informative about speaker's affective states, intentions, emotions, identity, health, and the relationships with the listeners. This para-linguistic information is usually not as precise in a written script. On the other side, text information in book reading, which is perceived with eye rather than ear, can be scanned in two dimensions (from top to bottom and from right to left). It allows reader to process content with more considering and an analytic view (see (Campbell, 2007)).

There are also some differences between reading a written script and listening to an acoustic signal. These perceptual differences are important when the objective is to generate an audio book. A reader can look at the previous and next words easily in text. So the information in a page can be considered like a stable picture without time passing conception. Unlike text, speech is heard in time. This difference provides an advantage for speech to play with its frequency, speech rate, tone, amplitude, stress, etc. But listeners have to receive and process more information online.

1.1.2 Expressiveness

Expressive speech synthesis deals with embedding various expressions related to different emotions and speaking styles in synthesis speech system. The emotion

and expressiveness are interchangeably used. Synthesizing expressive speech means being able to add thoughts, feelings, and emotions to words, sentences, and voices in a way that makes sense. So it can be inferred that expressive speech becomes richer than pure semantic information.

The relation between expressiveness and affects has been investigated in (Campbell, 2008), where it has been noted that the expressive speech exposes affects. In (Campbell, 2003), author had paid attention to non-verbal content such as non-lexical noises as important tool to express complex attitudes and intonations.

The term of expressiveness in speech synthesis is used to discriminate from *neutral* speech. Besides, since in speech synthesis, naturalness (human-like) is one on the main quality factor, an ideal synthetic speech signal should be expressive (in emotional situations) too. In other words, in speech quality, naturalness means not machine voice and expressiveness shows how much the system is able to synthesize speech with prosodic variations.

1.1.3 Speech quality

Based on (Campbell, 2007), **Speech quality** is a complex psychoacoustic outcome of the human perception process. In address to evaluate speech quality in speech synthesis context, there are many different ways such as diagnostics or comparative, subjective or objective, modular or global, task-based or generic, etc.

Generally speaking, speech quality has many perceptual dimensions. Commonly used dimensions are intelligibility, naturalness, clarity, pleasantness, brightness, loudness, listening effort. In (Kondo, 2012) speech quality is divided in two main aspects; the perceived overall speech quality and the speech intelligibility, whereas (Hinterleitner et al., 2011) underlines that speech quality needs to consider many different aspects such as overall impression, voice pleasantness, accentuation, listening effort, comprehension problems, acceptance, speech pauses (punctuation mark), intonation and emotion. An investigation has been done in (Hinterleitner, 2017) to find out the impact of these aspects on overall quality.

In speech processing, quality is a concept that should be defined with taking into account its context. In telecommunication applications, for instance, degradation factors such as acoustic noise, packet loss, or circuit noise can cause a decrease in speech quality and subsequently increase listening difficulty (see (Grancharov et al., 2008)).

The main problem of quality assessment is to find common definition of quality dimensions in a computational approach. In the following paragraph, two indicators of speech quality (intelligibility and naturalness) will be shortly considered. Afterwards, two practical methods will be mentioned which deal with quality definition.

The speech content should be understood. The *Speech intelligibility* is defined to measure how comprehensible speech is in special conditions. Intelligibility has been widely used to evaluate building or room acoustics, hearing aid performance, speech synthesis performance, and many others. According to (Ullmann et al., 2015), intelligibility is directly correlated with word recall statistics in utterance verification problem.

Speech naturalness is a term defined from the listener's perspective as how speech seems human-like? It helps listeners to get the speech message instead of focusing on the speech pattern. Based on (Edge, 2012), natural speech has also been defined as «typical speech you would expect to hear in any given situation».

As it is mentioned in (Kondo, 2012), the most reliable methods for speech quality evaluation are *subjective quality measures* that are based on the subjective opinion of listeners on the quality of the speech sample. The other alternative solutions for quality assessment are *objective evaluations*. While subjective assessment uses listeners to rate audible speech in terms of quality, objective assessment tries to implement algorithms for an automatic approximation of perceptual rating.

In this section, we have given a description of terms that will be used next to achieve common definitions. In the next section, the state of the art in speech synthesis will be considered.

1.2 Speech Synthesis

Over the last decades, TTS systems are rapidly developed concurrently with the growing of technology. Moreover the applications of TTS have been expanded. Speech synthesis as a crucial part of human-computer interaction encourages to improve the speech quality. While the synthetic speech quality depends on its application, some aspects of the quality like intelligibility has been attained to a fairly acceptable level. However the naturalness, expressiveness, and prosodic field in TTS systems need to be improved.

Technically, a TTS system could be divided in two parts: a front-end and a back-end part. The first part converts script to a linguistic specification and the second part uses

that specification to generate a waveform. The main differences between the most common TTS systems are in the back-end part. There are two basic types of TTS: the statistical parametric TTS and the unit selection based TTS. In the following, these two types and a hybrid approach will be reviewed.

1.2.1 Statistical Parametric Text to Speech Synthesis

The Statistical Parametric Speech Synthesis (SPSS) uses parameters instead of a corpus of stored speech units. Usually a trained acoustic model is used to predict parameters. These parameters feed a vocoder to generate the signals.

For a while, hidden Markov models (HMMs) had dominated acoustic modelling. The potential and flexibility of neural network lead the research on SPSS to use neural networks as the acoustic model. In (Zen et al., 2013), Deep Neural Network (DNN) has been introduced in SPSS. Several studies (Koriyama et al., 2015; X. Wang et al., 2016; Watts et al., 2016) compared HMM based TTS and DNN based one and confirmed the performance of these new models.

Recently numerous neural network based TTS systems have emerged. WaveNet proposed in (Oord et al., 2016), is a deep generative model of raw audio wave-forms which uses linguistic features, predicted log fundamental frequency (f_0), and phoneme duration as the inputs. Although WaveNet succeeds to synthesise human like voice, it suffers from high computational time. In (Sotelo et al., 2017), Char2Wav is presented as an end-to-end model that can be trained on characters. It is composed of an attention based, which is introduced in (Vaswani et al., 2017), auto-encoder as acoustic model and a neural vocoder. While Char2Wav relies on vocoder features from the WORLD TTS system (see (Morise et al., 2016)), in (Arik et al., 2017) new system called DeepVoice replaces all components with neural network. Tacotron has been introduced by (Y. Wang et al., 2017), as an end-to-end generative text-to-speech model that synthesizes speech directly from characters. This model has been improved as Tacotron2 in (Shen et al., 2018). In a new version of Tacotron, a sequence-to-sequence Tacotron-style model that generates mel spectrograms has been used followed by a modified WaveNet vocoder. An open-source toolkit named ESPnet-TTS has been introduced in (Hayashi et al., 2020; Watanabe et al., 2018) which supports state-of-the-art end-to-end TTS models.

As another approach, a Generative Adversarial Network (GAN) based TTS system

has been presented in (Saito et al., 2017). In this system, the acoustic model is trained to deceive the discriminator that distinguishes natural and synthetic speech.

1.2.2 Unit Selection Text to Speech Synthesis

The first attempt to build concatenative TTS was v-talk in ATR project (see (Sagisaka et al., 1992)). In (Black et al., 1995; Hunt et al., 1996), unit selection in concatenative speech synthesis was formalized as a optimization problem to find best candidate with the lowest cost.

The basic idea of unit selection (or concatenation) based TTS is to choose and concatenate a sequence of units such as diphone like in (François et al., 2001) or triphone in (Isogai et al., 2010) from a natural speech corpus. The selected units should minimize a cost function composed by the target cost and the concatenation cost (for instance a weighted sum of both). The target cost indicates the closeness between the linguistic features of the selected units and the target ones while the concatenation cost measures the differences of consecutive selected unit signals in their joins. Then the selection of units is a compromise between the minimization of the target cost and the concatenation one.

For example authors of (Toda, Kawai, & Tsuzaki, 2004) defined several sub-cost functions such as f_0, duration, and spectrum as target cost and f_0, phonetic category, spectrum as concatenation cost. They separately perform perceptual evaluations for optimizing each individual sub-cost independently and for optimizing the weights.

In addition to an optimization solution in unit selection engine, defining a long unit, for example syllable instead of phone, in corpus would reduce the concatenation cost. But the selection of long units generally require a larger voice corpus in order to cover varieties of units. Clustering units is another approach in concatenation based TTS. Authors of (Black et al., 1997) proposed to cluster units based on their phonetic and prosodic context before the selection process.

Figure 1.1 describes a comparable view of two basic TTS system procedures. SPSS is known for the smoothness of its generated signals and its flexibility to change voice characteristics and the prosodic feature. Conversely, unit selection based TTS systems provide more natural-sounding signals than SPSS (see (King et al., 2017; Zen et al., 2009)).

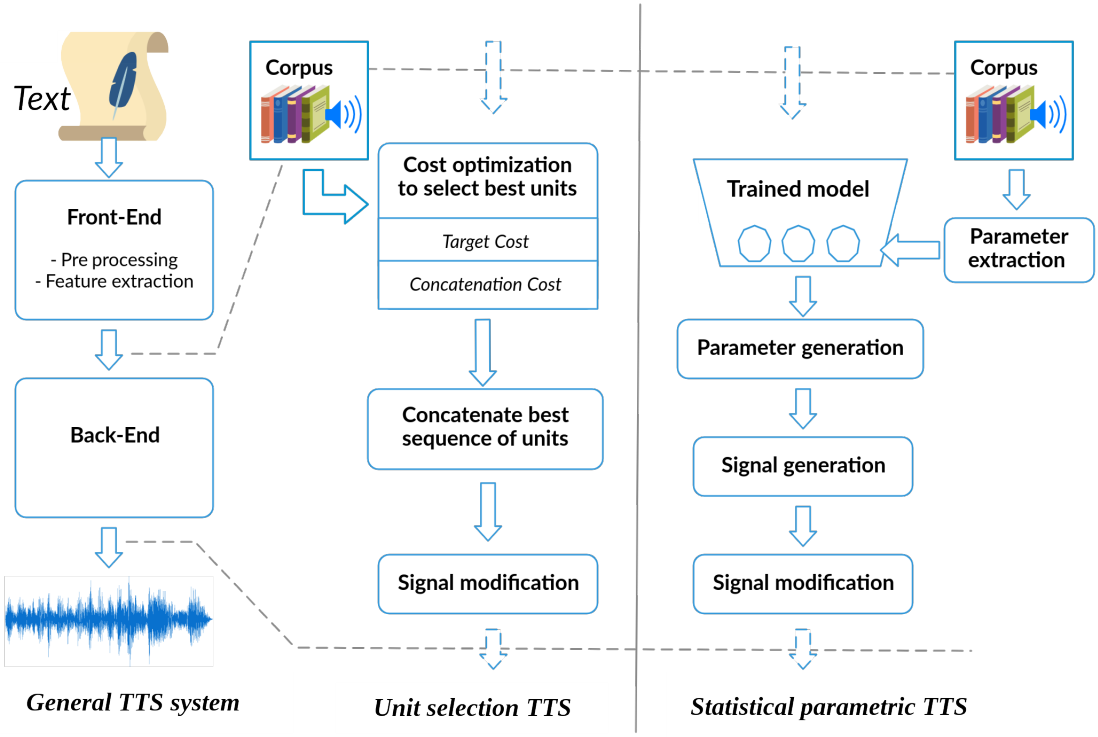


Figure 1.1 – A comparison between model based TTS and corpus based TTS and their components.

1.2.3 Hybrid Speech Synthesis

The advantages and disadvantages of two TTS types lead to the design of hybrid systems. The combination of both systems usually involves statistical models trained on the voice to predict parameters of an ideal generated speech and to guide a unit selection that concatenates real signal segments extracted from the voice. Recent studies and the last Blizzard challenges have revealed good achievements of hybrid systems (see for instance (Fan et al., 2014; King et al., 2018; King et al., 2017)).

The idea of hybrid TTS systems backs to HMM based TTS. For instance some studies like (Kawai et al., 2004; Rouibia et al., 2005) proposed to use acoustic parameters generated by a HTS for the target cost. Deep learning methods such as DNN and Recurrent Neural Networks (RNNs) have been successfully used as acoustic models in hybrid systems, replacing HMMs. In (Fernandez et al., 2015), it has been suggested to use a Deeply-stacked Bidirectional Recurrent Neural Networks (BiRNN) to deal with prosody cost within a unit selection TTS. But it was not used exclusively for synthesis-

ing.

Multisyn based on deep neural networks to guide unit selection systems has introduced in (Merritt et al., 2016). Results have shown using a DNN to generate features for calculating the target cost was more effective than using an HMM.

The main challenge in designing acoustic models is that the linguistic sequence does not have the same length as the acoustic sequence. For instance, in (Wan et al., 2017), a one-to-many approach is followed to deal with this problem. A LSTM-based auto-encoder is employed and permits to generate a sequence of acoustic frames representative of the input phoneme. As another example, in (Zhou et al., 2018), each candidate phone unit is converted into a fix-length unit vector, called *Unit2Vec*, and DNNs are used as target and concatenation cost functions. In order to manage the variable sequence length problem, a similar process has been applied in (Perquin et al., 2018), a feed-forward DNN for a one-to-one approach models phoneme frames, based on frame position, and the euclidean distance in the embedding space is used as the TTS target cost function. This approach also provides better results than an expert tuned target cost.

1.3 Expressive Speech Synthesis

Regardless the speech synthesis techniques, expressive speech synthesis needs to consider other elements.Based on four speaking styles that investigated in (Avanzi et al., 2014), it can be concluded the speech domains play an important role in prosodic consideration in TTS. It means some expressive domains like novel audio-book generation should be considered differently than news reading. Moreover the evaluation of synthetic speech will be reasonably different in expressive speech. So in the following, the previous works on expressive speech synthesis will be reviewed.

In expressive synthetic speech, the variation within training data in SPSS or unit diversity in unit based corpus becomes more important. The effect of linguistic, phonetic and prosodic expressive variations on the perception of expressiveness have been considered in (Tahon et al., 2017). They have used three speech corpora with different levels of expressiveness to create TTS voices. By comparing six AB tests, they concluded that high quality synthetic samples make better perception of expressiveness. Also the perception of expressiveness mainly relies on the adequacy of phonetics and prosody.

Some other studies such as (Alain et al., 2017; Iida et al., 2003) proposed to define a *normality*/*expressivity* score to each word of the corpus used to build the TTS voice. In other words, the corpus could be categorized in different types and level of expressiveness. Clustering speech data has been investigated in several studies such as (Eyben et al., 2012; Jauk, 2017; Székely et al., 2011). For example in (Székely et al., 2011), a Self-Organising Feature Maps (SOFM) used for clustering the expressive speech styles. Or the authors of (Eyben et al., 2012) proposed to improve expressiveness by clustering audio book data in HMM based TTS.

Most of the time expressive voice data are provided by audio books like (Eyben et al., 2012; Jauk, 2017; Székely et al., 2012). But ready audio books are have not been recorded to be used as corpus. An alignment process in armature audio book recording data had to be done in (Székely et al., 2012). Sometimes misalignment between text and speech of publicly available audio book is problematic. The authors of (Braunschweiler et al., 2010) tried to delete the differences in text and speech like insertions, deletions and substitutions made by the speaker with the help of a lightly supervised recognition.

In terms of technical consideration for expressive speech synthesis there are several studies. For instance (Theune et al., 2006) proposed a set of prosodic rules for converting neutral speech into storytelling speech. Some of these studies are able to synthesise speech for a given emotion and speaking style, such as (Charfuelan et al., 2013) which used audio book data labelled according to voice styles to control expressiveness in terms of discrete emotions or emotion dimensions or (Akuzawa et al., 2018) which uses an auto-regressive speech synthesis model with VAE (Variational Auto-Encoder). Besides, other studies like (Stanton et al., 2018) proposed to predict speaking style from text alone and use it to improve the expressiveness of synthetic audio book in an augmented version of Tacotron (Y. Wang et al., 2017).

Recently, in (Y.-J. Zhang et al., 2019), a VAE has been introduced to Tacotron2 (see (Shen et al., 2018)) to build an end-to-end expressive TTS system.

1.4 Evaluation

The success of any new technology largely depends on user perception of quality. Evaluation or comparing synthetic voices is necessary. Defining a precise platform and protocol provides the opportunity of the assessment. An investigation in perceptual

quality of speech and the influence factors can be found in (Hinterleitner, 2017).

Some quality metrics of synthetic speech like intelligibility need to be evaluated in particular scale. Since today's TTS systems have reached an acceptable level of intelligibility, the rest of this section focuses on other aspects of quality like overall quality, expressivity, pleasantness.

Speech quality assessment contains two main approaches. The perceptual test (subjective evaluation) is based on collecting subjective opinions (votes) from human test subjects following standardized procedures as specified e.g. in (Recommendation, 1996). The other approach lies on objective measures and is called objective evaluation. According to (Holub et al., 2017), objective evaluation is replacing human test subjects with relevant signal processing procedures to evaluate the synthetic signal in an algorithmic way. Generally an objective metric is an estimation of an aspect of quality. A strong correlation of the objective measure and a perceptual evaluation is the biggest challenge.

There is also another approach which is a compromise of perceptual test and objective evaluation. Evaluation modeling is an effort to approximate the perceptual quality based on some objective measures. Usually in evaluation modeling, a regression model explicitly predicts the perceptual quality like MOS score of naturalness in (Guo et al., 2020; Yoshimura et al., 2016) or overal MOS score of synthetic audio book in (Norrenbrock et al., 2012), while the objective measures provide a value which can be used implicitly as quality value.

These two evaluation approaches will be reviewed below in more details. Afterward evaluation modeling will be discussed.

1.4.1 Subjective Evaluation

Running a perceptual test by asking listeners' opinion is the most reliable method to evaluate the quality of speech signals. The most widely used direct approach is the categorical judgmental type. The listeners rate the quality of the test signal using a numerical scale. The measured quality of the test signal is obtained by averaging the scores of all listeners. It referred to as the MOS (see (Rothauser, 1969)). This method is one of the methods recommended by the IEEE subcommittee on subjective methods.

Another commonly used evaluation method is AB preference test. In this scenario,

a couple of signals, usually produced by two different systems, are presented to annotators who indicate their preference according to a given criterion.

Recently many studies use the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) listening test which is described in (Recommendation, 2003) . In this paradigm the listeners are asked to rate several systems between 0 and 100 (for naturalness from completely unnatural to completely natural). MUSHRA is effective at relative differences between multiple systems because listeners have knowledge of the full range of those systems before making they judgments. The main disadvantage of this test is its cognitive load. The MUSHRA could be exhausting specially when lots of evaluations have been asked and signals have a small difference.

In table 1.1, some of the most used subjective evaluation protocols are listed.

Test Name	Select/Grade	Reference	Detail
ABX	Selecting	Needed	Choose once which is more similar to the reference X
AB Preference	Selecting	Not Need	Force-choice paired
MOS	Grading	Just in Training	Evaluate from bad to excellent
CMOS	Grading	Not Need	Comparing score from much worse to much better
DMOS	Grading	Needed	Evaluate degradation from high to low
MUSHRA	Grading	Needed	Evaluate and compare MUlti Stimulus with Hidden Reference and Anchor
DAM	Grading	Not Need	Evaluate 16 measurements in 3 dimensions; parametric, metametric, and isometric

Table 1.1 – Subjective evaluation protocols

There are two main types of subjective assessment of speech quality. The signal quality can be requested in relation to a reference signal or evaluated without reference. By providing the reference signal, the listener opinion will be adjust to a baseline or the best point. For example in ABX test a reference is needed to evaluate the similarity, but the preference between two signals could be asked without reference.

Since the perceptual test is costly, its use is limited by the number of listeners and the number of samples. However, evaluating a small number of samples or using a small number of annotators prevents a generalization of the results and a good detec-

tion of small differences between two systems. Moreover when the number of listeners are not enough, the result could be biased by the small population. The effect of listeners number on the naturalness in MOS test is studied in (Wester et al., 2015). They concluded a stable level of significance will be only reached when more than 30 listeners are used. In (Chevelu, Lolive, et al., 2015), the selection of samples have been investigated. They suggested to select the most different samples in order to increase the significance of a perceptual evaluation. In comparison to a random selection method, their proposed method was more successful to distinguish quality in both HMM-based systems and unit selection systems.

As it is mentioned in (Loizou, 2011), although perceptual tests provide the most reliable method for assessment of speech quality, they are costly, time consuming and, furthermore, require to be done by trained listeners in most cases. This limitation lead researcher to find objective measures of speech quality at least to have an initial approximation of quality.

1.4.2 Objective Evaluation

Although an objective measure is not as accurate as perceptual test for quality evaluation, it does not need listeners. The objective evaluation provides an approximation of quality in shorter time and lower cost.

Objective quality measures can be classified based on the type of required information; *intrusive quality measures* need access to both the original and synthesized speech signal, while *non-intrusive quality measures* work only based on the synthesized signals. In order to evaluate quality degradation or acoustic similarity, waveform comparison algorithms are the main part of an intrusive objective quality measure. The timing misalignment of two signals is the main troublesome part in this kind of objective measure.

The Perceptual Evaluation of Speech Quality (PESQ) (see (Rix et al., 2001)) is a standard method to predict the quality of degraded signals based on its reference. This method was selected as the ITU-T recommendation P.862 (see (P.862, 2001)). A high correlation between PESQ and speech quality in telecommunication has been reported in (Hu et al., 2008; Loizou, 2011; Rix et al., 2001). As noticed by (Holub et al., 2017), the objective measures should be designed in regard to the application, language and speech quality type which is the prior of evaluation. The PESQ and most of intrusive

quality measures are efficient in telecommunication domain. Although (Loizou, 2011) noted however that, to some extent, speech intelligibility could be assessed by PESQ.

Mel-cepstral distortion (MCD) is one of the common intrusive quality measures used in speech synthesis. The MCD is calculated as an approximate log spectral distance between the synthetic signal and its reference (see (Prahallad et al., 2010)). MCD is used as objective measure of speech quality in speech coding, voice conversion, and SPSS for instance in (Luong et al., 2017; Perquin et al., 2018; Prahallad et al., 2010; Wu et al., 2016). It has been studied in (Toda, Black, et al., 2004) that as the MCD decreases, the corresponding voice quality is found to be better. Some studies such as (Perquin et al., 2018; Wu et al., 2016) used MCD as the cost function in acoustic models to be implemented in TTS system.

The main drawback of intrusive quality measures is that they need reference which is not available in most of TTS tasks. In addition to the intrusive objective measures, there are some reference-free objective measures to evaluate speech quality. The Kullback-Leibler Divergence (KLD) as a similarity metric between two distributions can be used as a non-intrusive objective measure.For example, in (Do et al., 2014) based on linguistic information and in (Ullmann et al., 2015) based on acoustic information employing KLD is proposed to estimate intelligibility. While methods used in (Do et al., 2014; Le Maguer et al., 2013; Ullmann et al., 2015) is HMM based TTS, the TTS cost of unit selection system can represent an objective measure of synthetic quality such as in (Chu et al., 2001; Krul et al., 2007; Toda et al., 2006). The concatenation cost can be an estimation of smoothness in unit selection TTS.

Based on (Loizou, 2011), an efficient objective measure should include a lot of information such as prosodics, semantics, linguistics and even psychoacoustics. In the following section the evaluation modeling will be discussed. In this approach, the objective measure is prediction model of perceptual quality.

1.4.3 Evaluation Modeling

In order to strike a balance between the reliability of perceptual test and the efficiency of objective evaluation, speech quality modeling is an alternative solution. The speech quality evaluation modeling can be described as a regressor or a classifier that try to find a relation between acoustic features and subjective scores. The ANIQUE which presented in (auditory non-intrusive quality estimation) model, (Kim, 2005), was

one of the first attempts to predict subjective quality of speech signal in telecommunication. In same time and same field, (Falk et al., 2006) focused on the noise and discontinuities to predict MOS score of corrupted speech.

The main challenge in this field is the feature selection. (Norrenbrock et al., 2012) reported a high correlation between an estimated MOS by using MFCC and prosodic features and a subjective MOS. While (Li et al., 2014) proposed to use Gabor filter bank to extract high dimensional spectrotemporal features, (Dubey et al., 2015) used multi-resolution auditory model which simultaneously takes into account frequency and time domain. Recently (Hakami et al., 2017) have shown that an augmented feature set can reduce the effect of noise. The proposed auto-encoder and a linear regressor as a neural network model help to improve prediction of the quality.

(Norrenbrock et al., 2012) warned that a joint research on the feature and the model level is necessary. The ANIQUE, presented in (Kim, 2005), is based on the temporal envelope representation of speech. (Li et al., 2014; Norrenbrock et al., 2015) reported encouraging results with Support Vector Regression (SVR) in deal with quality assessment metric for enhanced speech signals. (T. Zhang et al., 2016) claimed the SVR has two main drawbacks; firstly, it needs expensive tasks to labeling and features extraction, secondly the labeling results are mainly coming from person's subjective feeling. In (Fu et al., 2018), a Bidirectional Long Short-Term Memory (BLSTM) model is implemented to predict the utterance-level quality. Since this model has been designed for telecommunication tasks, it had been found a high correlation with PESQ score.

Most of the previous works on evaluation modeling has been done in telecommunication framework rather than speech synthesis. There are also some studies that predict perceptual quality of synthesized speech for example (Guo et al., 2020; Norrenbrock et al., 2012; Yoshimura et al., 2016). In (Yoshimura et al., 2016), a training based on subjective evaluation results is proposed. This study revealed that a Convolutional Neural Network (CNN) by identifying local signal features could improve the prediction of naturalness in synthetic speech. As another methodology in evaluation modeling, a residual learning network shows a good performance in predicting naturalness by (Guo et al., 2020).

In order to evaluate the performance of these models, usually the Pearson correlation and Root Mean Squared Error between estimated MOS and subjective MOS are calculated. We can also notice that Spearman rank order correlation coefficient is used by (Li et al., 2014) and a MOS-based rank relevance is calculated in (T. Zhang

et al., 2016). This means that pairwise comparison of systems are more important than trusting the resulted value of models.

Finally (Gupta et al., 2017) can be mentioned as a new method in speech quality estimation. Authors have explored use of two neuroimaging techniques (EEG and fNIRS) to better understand neuronal and cerebral haemodynamic changes resulting from synthesized speech of varying quality. They tried to model neuronal and physiological measures (e.g., heart rate changes) as perception quality of listener.

1.5 Conclusion

The framework of the thesis has been described in this chapter. Some terms have been discriminated and described. The state of the art TTS systems have been introduced and their main differences have been highlighted. Finally the measures and the protocols which can be used in evaluation of synthetic speech have been reviewed.

The rest of this study will be concentrated on audio-book generation which is an expressive speech synthesis task. The main perceptual quality aspect will be the preference of the listeners when they listen to the signals.

We will use unit selection based TTS in this thesis. Although, in TTS, vocoder-based approaches – like end-to-end DNN systems – are more and more prevalent, hybrid or classical unit selection-based systems are still well-adapted to take into account the data parsimony constraint. However, their achievements are very sensible to the voice quality and the impact of the voice is all the stronger as the constraint of parsimony is important (see (Chevelu & Lolive, 2015; Lambert et al., 2007; Szklanny et al., 2017)). Also this kind of the TTS provides an comparable objective measure such as concatenation and target cost which will be useful for evaluation of signals.

CHAPTER 2

CORPUS DESIGN

A TTS system needs a voice corpus which is basically an aligned set of texts and speech signals. Unit selection based TTS system uses this corpus as a unit data-base while the SPSS system uses it for its training. In practice, the synthetic speech quality is strongly affected by the quality of the voice corpus. Previous studies (Bozkurt et al., 2003; Chevelu & Lolive, 2015; Isogai et al., 2010; Lambert et al., 2007) showed the importance of the content of voice corpus. This is true especially for unit selection-based speech synthesis but also statistical parametric and hybrid ones. Due to the natural heavy-tailed distribution of linguistic events, a random (or unbalanced) voice corpus contains lots of unit repetitions and, most importantly, it does not guarantee a sufficient variety of units for the speech synthesis process.

There are two main approaches to prepare TTS voice corpus with a size constraint, and this constraint can have several motivations, economic or technical. Sometimes a ready voice corpus should be reduced in order to increase the adequacy of voice to the application context. In particular, in case of unit selection based TTS systems or hybrid ones, a reduced voice corpus size may also accelerate the synthesis process considering the smaller search space. On the other hand, the voice corpus design can be processed by selecting a text script whose vocal reading will be recorded. In this approach, the script should be as small as possible in order to minimize the human cost of high quality recording and labeling processes. To sum up, a well-designed corpus combines parsimony and balanced unit coverage in order to gain a satisfactory level of richness with a minimal cost construction. These two approaches and their motivation will be detailed in sections 2.2 and 2.3.

This chapter explains the problem of the voice corpus design. By reviewing the literature, the corpus design will be surveyed in voice corpus reduction and text selection sections.

2.1 Audio-book generation problem

According to (Van Santen et al., 1997) the speech corpus design can be formulated as the following optimization problem: *"selecting the shortest subset of sentences from a huge corpus in the way that the subset offers a balanced phonetic and prosodic coverage. The corpus could be text before recording or speech signal"*. Moreover, the main objective of a TTS system is to generate speech signals with the highest possible quality. The richness of voice corpus, measured, among others, by rates of phonetic and prosodic coverage, influences this quality, but finally a cognitive aspect like perceptual quality could be a criterion to assess the achievements of a voice corpus in the TTS framework. Although, in case of building a parsimonious voice, coverage of all possible units may not be possible, having the best (partial) coverage for a given size is crucial for TTS.

Beside reducing the cost of corpus preparation by optimizing the corpus contents, speech corpus design could be used for creating TTS data-base in low resource languages such as for Spanish (Umbert et al., 2006), Basque (Saratxaga et al., 2006), Catalan (Bonafonte et al., 2008), Arabic (Halabi et al., 2016), or South African languages (Van Niekerk et al., 2017).

The selection of a minimal sized subset from a sentence set under the constraint of covering a given number of linguistic units can be formalized as a set covering problem (François et al., 2001). This set covering problem is an optimization problem, which is NP hard (Karp, 1972).

Due to this complexity, heuristic approaches are necessary to solve the problem on large databases with a reasonable computational time. In the following paragraphs, the main algorithms that are used for this problem will be reviewed.

Different approaches for finding an approximated solution for this NP-hard problem are possible. The most commonly used algorithmic strategy is the greedy based algorithm (Barbot et al., 2015; François et al., 2002; Ni et al., 2006).

The greedy optimization is of course sub-optimal but offers advantages. Firstly, it is computationally very efficient. Secondly, it provides a local optimal solution.

The greedy algorithm can be categorized in two main classes. The first one, called agglomerative greedy, which begins with an empty set, works with an iterative selection of sentences among a large sentence set to build a reduced and rich corpus. The second one, called spitting greedy, which starts with the whole sentence set and reduces

it by removing the less useful sentences usually composed of redundant units. The performance of these two approaches (and a pair exchange method) have been investigated by (François et al., 2002) in the TTS framework. It has been shown that the pair exchange method does not guarantee a total covering of the given set of attributes, contrary to the greedy methods. The agglomerative greedy was just slightly better in terms of solution size and computational time in comparison with the spitting greedy and the pair exchange method. The combination of these methods has demonstrated that it could be more efficient if the spitting greedy is applied after a agglomerative greedy. The closeness between a greedy solution and optimal one has been shown in (Barbot et al., 2015) for a given sentence score and an agglomerative-spitting greedy strategy.

Others approaches using the greedy strategy also exist like a weighted greedy algorithm based on the unit frequency that has been proposed by (W. Zhang et al., 2010) for the set covering problem in TTS voice corpus.

As an alternative to the greedy algorithm, Lagrangian relaxation principles permit to reach an optimal solution for set covering problem in case of a small search space. In (Barbot et al., 2015), the performances of a Lagrangian relaxation based–algorithm are compared with the ones of a greedy-type algorithm (combination of the agglomerative and spitting strategies) to extract a sentence subset from a large corpus. Even if the Lagrangian relaxation based algorithm gives better solutions, it also provides a lower bound of the minimal size of a unit covering that permits to observe the good closeness of solutions derived by the greedy-type algorithm to the optimal one with a smaller computational cost. Therefore, authors conclude the greedy approach is the most adequate strategy.

A different approach has been tested in (Espinosa et al., 2010) using machine learning techniques for this problem. The authors proposed to train a SVM that is able to predict a ranking of *utterance utility*. The *utility* of an utterance is calculated by synthesizing a target utterance set with the voice under construction, summing the corresponding concatenation and target costs, and then adding the utterance to the voice and synthesizing the target set again. The global cost difference between these two syntheses is the utility of the added utterance. Some features corresponding to each utterance, each voice corpus and the target script which is supposed to be synthesised have been extracted. The SVM has been trained to rank utterances based on their features (such as number of diphones and words, rareness or commonness of diphones,

etc.) in order to drive the selection of utterances to be added to the TTS voice corpus.

In our approach, in order to reduce the audio-book generation cost, the book script will be divided into two parts: a recorded one and a synthetic one. The recorded part is composed of natural speech signals and will be used as the TTS voice corpus for synthesising the other part. In comparison with previous studies, this selection would be more complex as the order of synthetic and recorded utterances in the final audio book could impact the overall perceived quality of the final audio book.

Letting the target text, which is supposed to be synthesised, to be changed emerges the complexity of problem in comparison with a TTS system for general domain. It means considering the synthetic part could also help the voice corpus design process to improve the final quality of audio book.

The problem of generating an audio book can be formulated as follows. Let us define the book as a list of utterances $U = (u_0, u_1, ..., u_n)$, where u_i is the i^{th} utterance in book. A subset of U should be selected as the recording script. This selection can be described by the vector $R = (r_0, r_1, ..., r_n)$, where $r_i \in \{0, 1\}$ and $r_i = 1$ means the i^{th} utterance is in the recording script. The recording cost $C(U, R)$ is basically defined as the total length of the selected utterances which compose the recording part. The quality of the final audio book is noted as $Q(U, R)$. Consequently, the optimization problem can be formulated as:

$$\arg \max_{\substack{R \in \{0,1\}^n \\ C(U,R) \leq l}} Q(U, R)\,.$$

In other words, the problem is to find which R would provide the maximum quality of the final audio book with respect to a maximum length l of the recording script.

Regardless of the optimization problem above, a corpus used for the text-to-speech system generally under parsimony constraint, while guaranteeing voice quality, can be prepared in two ways. These two strategies and their related previous works will be reviewed in the following sections.

2.2 Speech corpus reduction

In this approach a recorded voice is ready but it should be pruned, optimized, and reduced. The main reasons of corpus reduction are memory limitation, labeling cost or existence of some destructive data. Adding to these, in the case of TTS approaches

based on concatenation, a smaller corpus helps unit selection to speed up. In other words, the selection of units for concatenation could be pre-selected by corpus design.

Several studies can be found in literature which deal with this kind of problem. The investigated approach in (Krul et al., 2007) is to delete the part of the voice corpus least used when synthesizing a test set of domain specific utterances, in order to decrease the voice corpus size. This result has been compared with the voice provided by a KLD based reduction approach. This strategy consisted in iteratively selecting phrases (breath groups) such as the distribution of the voice under construction has the unit distribution closest to a target one. The results indicated a better performance for a domain-based pruning method in comparison with KLD-based methods.

Some studies have been done in order to use audio book data for general-domain TTS systems. The main goal, in this case, would be to produce a more natural signal by extracting neutral voice part of audio book. For instance, an outlier-removal approach has been used by (Braunschweiler et al., 2011; Cooper et al., 2016) for HMM-based TTS framework. The outlier, which causes less natural-sounding voice, is founded out as hypo-articulated utterances and utterances with a low mean F_0 in (Cooper et al., 2016). A preliminary study in (Braunschweiler et al., 2011) has shown that a HMM-based TTS gives a better perceived voice when the non-neutral style sentences are removed from the learning corpus. In this study, authors discarded sentences based on acoustic features (extreme F_0 patterns, too loud or barely audible sentences) and linguistic features (non-neutral style sentences such as quotation, interjections, utterances starting with lowercase , etc.). Sometimes, in found data like available audio books in the public domain, some destructive parts can be discriminated. It means selecting a smaller, cleaner subset for voice is better and less time consuming than building from the full noisy data-set. In (Baljekar et al., 2016), two types of errors have been discovered to be removed as misalignment and annotation errors due to noisiness of signals.

(Chalamandaris et al., 2014) have compared synthetic signals from three TTS systems, each one using a different corpus. The first system has been developed based on the entire audio data-set without any pruning. The second one used pruned data based on prosodic features (mean and standard deviation of the pitch value) at the phrase level. The corpus of third TTS system has been obtained by pruning using prosodic features and a segmental criterion. The latter criterion was how appropriately is an aligned sentence annotated. The listener preference indicated that although a

simple prosodic pruning does not help to improve significantly the quality of synthesis, a pruning based on both prosodic and segmental features leads to better synthetic speech. The F_0 variations as some prosodic features also have been taken into account in (Isogai et al., 2010) in the coverage metric, besides coverage of syllables, for voice corpus reduction.

To finish this section, we also have to mention that some works on voice corpus design also exist in speech recognition. While the task is different from speech synthesis, studies like (Itoh et al., 2012; Shinohara, 2014) still focus on the distribution of units as database. In (Shinohara, 2014), it is suggested to select phonetically-balanced sentences and, in (Itoh et al., 2012), it is proposed to use an entropy-based method for training data selection.

2.3 Selecting text to record

The main difference between text selection and voice reduction is the available information. Acoustic information is absent in text selection problem. Usually in this kind of problem, text processing techniques are more used than signal processing techniques. The extraction of features from text should also be adapted to voice corpus design in the TTS framework. Moreover, there is less information (for example related to expressiveness, speaking style and intonation) in text than voice to use it in text selection process for voice corpus design. An ideal end-to-end TTS should infer some information such as prosody from text directly. This inference can be used in text selection, however there are usually some errors.

Logically, the selected script should be as small as possible while it should be both phonetically and prosodically rich. Selecting an efficient script before recording contributes to reducing storage memory, recording and annotation costs.

Several works on automatic TTS corpus design have been carried out since early 2000s (for instance (François et al., 2001; Gauvain et al., 1990; Kawai et al., 2000; Van Santen et al., 1997) for some preliminary ones). The feature extraction part is the initial part which provides information for the selection algorithm. The linguistic features in different levels were used in previous works. They could be diphone (François et al., 2001) or triphone (Isogai et al., 2010) labels, phonetic "sandwiches" (Cadic et al., 2010), etc. Some may add some positional characteristics to these units (Chevelu & Lolive, 2015) or some stress information (Lambert et al., 2007).

Regardless of the definition of unit in the feature extraction step, there are two main approaches for selection: a set of defined units could be covered by the selected part, or the selected part could respect the original or a target unit distribution. Some studies also tried to combine both ideas. For instance, in (W. Zhang et al., 2010), partial coverings of diphones (the best obtained coverage rate was 93,52% of diphones) were derived using a greedy algorithm where the score of each utterance depended on the frequency of its units in the initial corpus (rare units had higher weights than more frequent units): "*applying high weights for the rare units can improve the performance in the situations that complete coverage is feasible*". These two ideas are illustrated by figure 2.1. For this illustration, three types of elements are considered (triangles, circles and stars) and the selection size is limited to three elements. The coverage-based approach selects a representative element (highlighted in red color) in each category, whereas the distribution based one selects elements (in blue) to match the original type distribution. The previous works which follow these approaches will be detailed next.

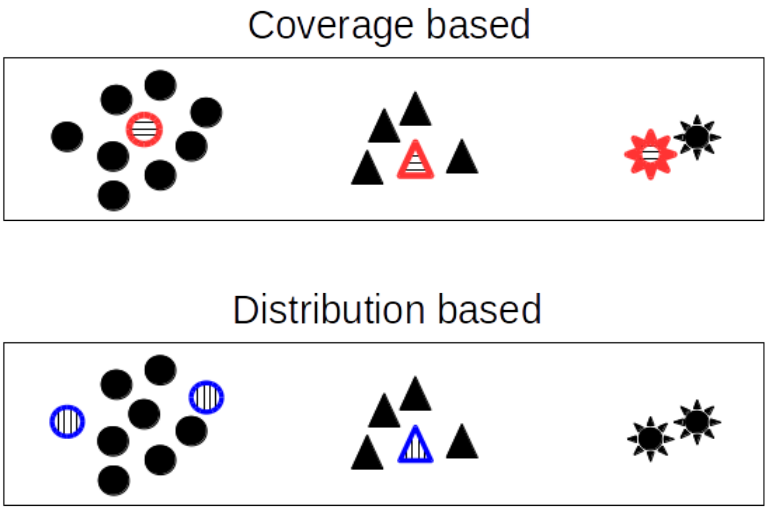


Figure 2.1 – Selection of three elements out of sixteen elements of three types by covering-based versus distribution-based approach

2.3.1 Covering-based approaches

Some preliminary studies on text selection for speech processing dealt with the maximization of linguistic unit covering (usually phonemes, diphonemes and triphone-

mes) as described in (François et al., 2001; Gauvain et al., 1990; Van Santen et al., 1997). The covering of linguistic units under a parsimony constraint has been the main idea of script corpus design. The utterance selection as the set covering problem for TTS voice corpora has been formulated in (François et al., 2001).

Unit selection based TTS systems select units from the voice corpus based on their similarity with target units. Similarity is evaluated using a target cost. If this similarity measure consists on a binary value (*match/not-match*), the covering of all units becomes crucial. Otherwise, unit selection based TTS systems would fail to synthesize some given text.

Selecting the utterances by identifying clusters of acoustically similar units at synthesis time was the idea considered in (Black et al., 2001) for script selection.

Introducing a different strategy to build a linguistic covering, authors of (Cadic et al., 2010) focused on sentence construction instead of sentence selection from an initial corpus in order to maximize the vocalic sandwiches covering rate (VSCR). Indeed, they used symbolic features such as phonetic/linguistic context to approximate the concatenation cost. Based on the final quality of their TTS system, they concluded that vocalic sandwiches are more suitable than traditional units. Their experiments showed that the best process to construct sentence was not completely automatic and required human supervision time. Loosing semantic coherence in built sentences is the main flaw of the method they suggest. In addition to this human cost, one of the main drawbacks of this method they suggest is the loss of semantic coherence in the constructed sentences. Moreover, the final corpus does follow a coherent prosody in an expressive context.

Regardless of defined units to cover (such as diphonemes, triphonemes or POS (part of speech) tags in (Barbot et al., 2015; Ni et al., 2006)), many strategies can be used to select an utterance subset offering a rich linguistic variety. For instance, in (Barbot et al., 2015), the Lagrangian relaxation based algorithm computes total n-coverings (utterance sets covering at least n times each unit) which are refined using heuristics based on Lagrangian relaxtion principles. Similarly, a spitting greedy algorithm begins with a total covering (corresponding to the whole initial corpus), whereas an agglomerative greedy one iteratively complements a partial covering until reaching a stopping condition. In order to reach as close as possible a n-covering under a constraint of size, others approaches are also possible: as illustration, one can start to cover each unit at least once and afterward this 1-covering is complemented by an iterative incrementation of the number of required unit instances (Shamsi et al., 2019a), another

one may consist on covering basic elements like phonemes first and after considering larger and larger units such as diphonemes, triphonemes, etc.

2.3.2 Distribution-based approaches

Besides the covering approach, some studies (Krul et al., 2006; Nose et al., 2017; Saratxaga et al., 2006; Van Niekerk et al., 2017) investigated the use of distribution of units in the corpus. Generally, it turns out that the target distribution is the natural or domain-specific ones, favoring the presence of several instances of common units in the covering. For example, covering several times a common triphone provides more prosodic diversity that covering a rare triphone, and should improve the quality of synthetic signals. This is true especially for a defined target script which is supposed to be synthesized. The unit distribution of the defined target script could be different from the general distribution of units in language. In an exceptional situation, coverage of a unit that does not exist in target script is not necessary.

The phonetic and lexical balance was the main concern in (Saratxaga et al., 2006). The purpose was to design a corpus with most similar appearance rate of units as their appearance in the language. (Krul et al., 2006) suggested to design TTS corpora in the way that the KLD between their diphoneme and triphoneme distribution and a prior distribution would be minimized. Recently (Nose et al., 2017; Nose et al., 2015a) have paid attention to different attributes of corpora for unit selection and statistic parametric TTS. They have proposed a sentence selection technique for constructing phonetically and prosodically balanced corpora, named extended entropy of phonetic and prosodic contexts. The experimental results demonstrated that the proposed method achieved better coverage and balance of both contexts in three languages (Japanese, English, and Chinese). It emphasizes the importance of prosody and contextual information in corpus design.

2.4 Conclusion

In the previous sections, the TTS corpus design problem has been explained. The technical solutions of sub-set selection have been reviewed. The thesis objective has also been formalized as audio-book generation problem by division of an audio book in two parts: a synthetic part and a recorded part which is used as TTS voice corpus to

synthesise first part. Afterward the previous works have been categorized into two main approaches. The speech corpus reduction approach which profits from acoustic information as well as linguistic information. And a text selection approach before recording process which is close to our original problem. The previous works on text selection based on coverage of units or distribution based methods have been presented.

The complexity of the problem leads us to simplify it. In the first step the impact of synthetic/recorded utterance order in final audio book will not be taken into account. This study will be focused on unit selection TTS corpus design for French audio-book generation. Finally, even if the original problem in this thesis is the text selection before the recording process, we use an already fully recorded audio book to simulate the selection process and to be able to evaluate perceptually different strategies.

CHAPTER 3

FROM CORPUS REDUCTION TO SCRIPT SELECTION

Script selection for voice corpus design should be revised based on the application of voice corpus. For example a voice corpus for command recognition systems would be different from a voice corpus for expressive audio-book generation task. In order to adapt the script selection to our TTS task, a voice corpus reduction process could be helpful to identify the required characteristics of optimal script portion to select.

In this chapter, the TTS voice corpus reduction problem will be investigated. The analysis of the best sub-set of a voice corpus could help to select a script for TTS voice corpus. It means a posterior strategy will be followed, requiring the estimation of the best solution when the fully recorded book is ready. Therefore, the goal will be finding a solution based on textual and linguistic information which is as close as possible to the voice corpus reduction result.

The corpus of the original recorded voice is an expressive high quality audio book which is spoken by professional speaker. So regardless of the quality of synthetic signals produced by TTS, it is assumed that corpus reduction causes quality degradation of the output signal sequence which contains natural recorded signals and synthesized ones.

This section is structured as follow. Section 3.1 will explain the challenge that is faced in TTS voice corpus reduction and the solution that will be employed. Afterward in section 3.2, the objective measures for synthetic quality assessment and the metrics that can be used for ranking candidates in the reduction process will be investigated as a preliminary experiment. Finally, section 3.3 will evaluate the proposed solution for voice corpus reduction by comparing it result with a random strategy result.

This work has been published as a conference paper in (Shamsi, 2020).

3.1 Optimization strategy

In the ideal scenario, for a voice corpus and a given reduction length, all possible subsets should be evaluated perceptually. Unfortunately, this is not possible in a reasonable time. In practice, in order to reduce the voice corpus, two main requirements are needed: a practical subset selection heuristic and an automatic evaluation method to assess the quality of synthetic signals based on a given subset. By considering the previous works (Barbot et al., 2015; Espinosa et al., 2010; François et al., 2002), the greedy algorithm had been found to be a practical solution to find a sub-optimal portion of a script in reasonable time.

In the greedy algorithm, whatever is spitting or agglomerative, a ranking measure is needed to evaluate candidates. In each step of spitting greedy process, the best candidate will be selected to be removed from the corpus (or to be added to the voice corpus in agglomerative greedy process). For each reduction rate, the rest of the book would be synthesized and evaluated in terms of quality.

A similar process has been implemented in (Espinosa et al., 2010). They proposed to agglomeratively select utterances which give results with the best TTS costs for synthesizing a set of utterances. In our problem, this set corresponds to the rest of the book made of the sentences not selected in the voice corpus. Measures for ranking utterances and evaluating the quality can be the same metric. Like in (Espinosa et al., 2010) which the TTS costs is used for these two purposes. They will be investigated in section 3.2. The selection algorithm and the computational problem will be investigated in the following.

The selection process starts with the full corpus as the voice corpus set (VC). In each step of spitting greedy, a portion of the voice corpus (one utterance) will be removed from VC. The selected utterance is added to the synthetic part set (SP). The IRISA TTS system (Alain et al., 2017), which is unit selection-based, uses VC for synthesising the rest of the book SP. In each step of this process, not only a larger part of audio book is being replaced with synthetic signals, but also the voice corpus that will be used by the TTS system becomes smaller. So a small change in voice corpus could effect on the final quality.

Some utterances in VC contain unique units and a concatenative TTS couldn't find these units in other utterances. These utterances will be locked (VL) and should not be removed from VC. The remaining utterances in VC which have not been added

into VL are the candidate set (CS) for next selection step. The spitting greedy process is continued until the CS becomes empty. This process can be described as the algorithm 1.

Algorithm 1: Spitting greedy for optimizing corpus reduction

```
VC = CS = all utterances ;
SP = VL = ∅ ;
while CS has at least one utterance do
    for All U_i utterance in CS do
        Remove U_i from VC and add to SP ;
        Synthesis SP by using VC ;
        if synthesizing of SP failed then
            Add Ui to VL and remove from CS ;
        else
            Evaluate synthetic signals of SP and save as quality reduction of U_i ;
        end
        Remove U_i from SP and add to VC ;
    end
    Find U_x as the minimum quality reduction from U_i in CS ;
    Remove U_x from VC and CS and add to SP ;
end
```

In order to simplify the evaluation of quality, the order and configuration of synthetic-recorded utterances in the final audio book will be ignored in this phase. It means the final audio book will be evaluated based on the quality of only synthetic part (SP).

The first challenge to use this algorithm is its computational complexity $O(n^3)$. Calling TTS for synthesizing SP and then evaluating its quality are expansive. In next subsections, ideas that can be applied to reduce the computational problem will be mentioned. Generally, we expect that speeding up the algorithm would reduce the optimality of solution.

Avoid repetitive calculations

In unit selection-based TTS, the output signal is generated by the concatenation of speech segments from the voice corpus. Storing a dependency matrix between the utterances in SP and the utterances in VC in each step can be used to avoid

the synthesis of all the utterances in SP in the next step, when only one utterance is removed from VC. It means removing one utterance (U_i) from VC will change only the synthetic signals of utterances in SP which used speech segments from U_i. The dependency matrix can be helpful to avoid repetitive synthesis and evaluation.

By applying this modification for two small corpora, the computational time is reduced by 67% and 71% respectively in the corpora with 168 and 334 utterances. Although this reduction of computational time would be more helpful in larger corpora and also does not impact on the final solution, it does not seem to be enough.

Merging utterances

Another idea which reduces the computational time is merging consecutive utterances. It helps to have less combinations and selection choices in CS. It seems rational and practical to ask the speaker to record multi-utterances or a paragraph instead of only one utterance. By following this strategy for a small voice corpus (334 utterances), the overall quality of the synthesized part is decreased drastically. In other words, by merging more utterances the estimation of overall quality (TTS cost) becomes more similar to random selection. The figure 3.1 shows the impact of merging utterances in reduction process by spitting greedy algorithm.

In figure 3.1 the TTS global cost is used as the metric to assess the quality degradation of the final audio book. It means, by reducing the voice corpus, the synthetic quality will be reduced. Based on this experiment, the reduction process by removing only one utterance (blue line) in each step performs better than random. But by removing a bunch of 5 and 10 utterances the quality degradation is equal or even higher than random strategy.

The gain of computational reduction, by merging utterances, is not worth the cost of loosing efficiency.

Corpus reduction based on the initial ranking

By investigating the ranking list of utterances in consecutive steps of the algorithm, it has been observed that a ranking list of candidate utterance does not change a lot from one iteration to the next. It means that the rank of utterances for removing from VC does not depend on the size and content of VC and SP. Although a simple experiment has showed that this assumption is not completely true, considering the rank of utter-

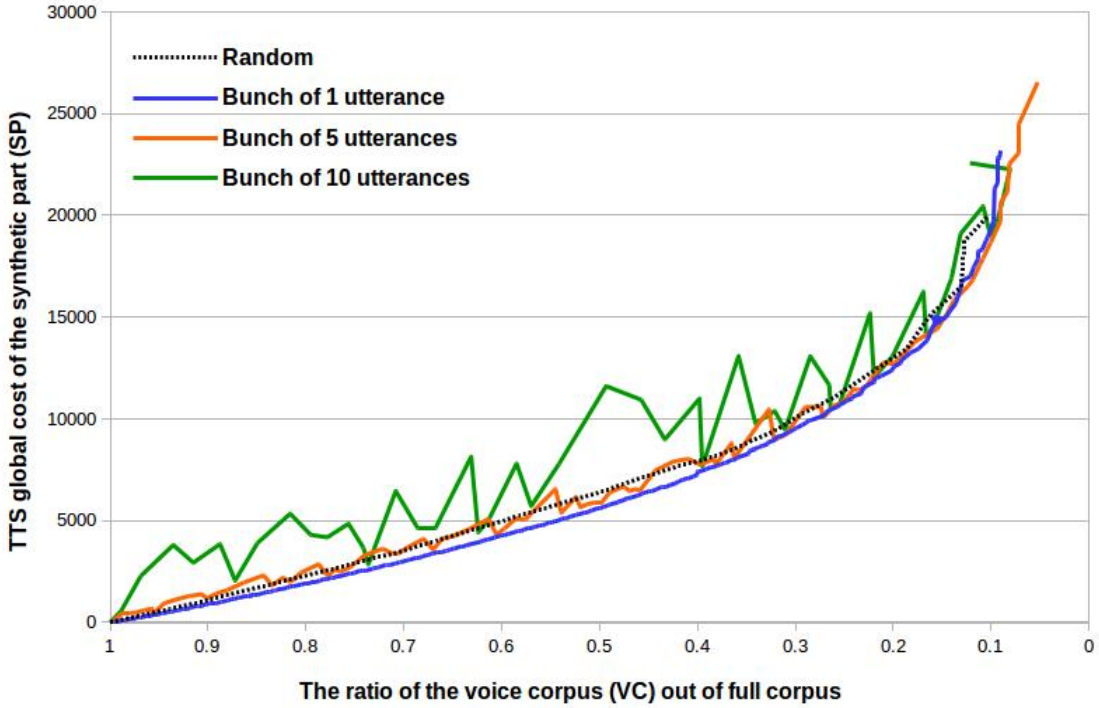


Figure 3.1 – The quality degradation of synthetic part (based on TTS global cost) in reduction process of a voice corpus with 334 utterances (1 hour) by merging utterances.

ances, for being in voice corpus, as a stable list helps to get rid of the computational problem. The time complexity of the algorithm considering this modification decreases to $O(n)$. The normalized root-mean-square error (NRMSE) of the TTS global cost for corpus reduction from the initial full corpus down to 70% of its size is 97 when this value is increased to 16727.It shows that, in a big corpus, keeping the initial ranking list makes the resultant sub-corpus slightly worse than following the original spitting greedy.

In the following some strategies closer to the optimal than following the initial ranking will be considered.

In order to have a trade-off between optimality and computational time, the assumption of the stability of ranking list can be used. Instead of transferring utterance by utterance from VC to SP, a bunch of utterances at the head of the ranking list can be removed from VC and added to SP. Moreover, contrary to the previous approach where the list is kept fixed once and for all, it can be updated from time to time. For instance, after several times that a bunch of utterances has been removed from VC, we

can consider updating the ranking list. It will help to complete the algorithm in shorter time.

Three methods to update the ranking list, which can work with different optimality and computational time, are proposed:

1. **Re-ranking head of candidates ranking list:** At each iteration just a bunch of utterances (100 utterances) would be removed so it seems to be a rational approximation that only head of the ranking list (200 utterances) is re-ranked.
2. **Re-ranking candidates n times in whole algorithm:** The other solution could be re-ranking the whole of candidates list but just in certain iterations (3 times; 0% of reduction rate, 33% of reduction rate, 66% of reduction rate).
3. **Re-ranking candidates independent of SP:** Indeed a small change inside the voice corpus VC could change the synthetic quality of all synthetic part SP. So for evaluation of U_i in the candidate list, the synthetic quality of SP, which U_i has been added to, should be considered. We propose to consider only the synthetic quality of U_i for its evaluation in the ranking list. The idea behind this proposition is that the small change of VC by removing U_i could be ignored. Only the quality degradation of the final audio book, because of replacing recorded voice U_i by its synthetic signal, would be taken into account. This idea changes the algorithm 1 by modifying SP to U_i for synthesis and evaluation (line 6-10).

Indeed skipping some computations in the greedy algorithm helps to find a solution in a shorter computational time. Nevertheless, the subset solution would be closer to the random solution and its goodness would be degraded.

The greedy strategy is employed to find subset solutions in reasonable time. The original greedy algorithm for selecting utterance by utterance in an audio book with thousands utterances is computationally expansive. It is estimated that if an atomic operation is the synthesis and evaluation of an utterance and each operation takes only one second (an optimistic estimation based on IRISA TTS and our facility), this experiment takes more than years to be completed. The table 3.2 compares methods with different approximation level in terms of computational time. Based on these estimation, approximation methods at level three seem to be more practical. While it is not obvious which method in this approximation level is more efficient for our problem, the method with the lowest computational time is selected for the experiments. It means the selection function of the greedy algorithm will evaluate candidates only based on their synthetic signal instead of the synthetic quality of SP.

Approximation level ↓

Level	Method	Num. Op.	Time
0	Consider all combinations	10^{10000}	-
1	✔ Remove utterance by utterance ✔ Update full ranking list at each step	10^{10}	169 Years
2	✔ Remove bunch of utterances (100) ✔ Update full ranking list at each step	61,899,200	716 Days
3	✔ Remove bunch of utterances (100) ✔ Update the head (200) of ranking list	10,560,00	122 Days
	✔ Remove bunch of utterances (100) ✔ Update ranking list just 3 times	4,958,415	More than 57 Days
	✔ Remove bunch of utterances (100) ✔ Update ranking list just based on candidate (synthetic part independent)	110,187	2 Days
4	✔ Remove bunch of utterances (100) ✔ Use initial ranking list	56,139	15 Hours

Figure 3.2 – Computational time estimation for different approximation levels using an audio book with 3339 utterances. Synthesis and evaluation time of each utterance has been estimated to be one second.

Regardless of the computational time problem, the greedy algorithm needs an objective metric for selection and an automatic measure for evaluating synthetic quality. While our final task is expressive audio-book generation, the selected sub-set as a corpus can be evaluated differently for example by concerning its linguistic features. As a preliminary experiment, these two objective metrics will be looking for in the following section. Needless to say that they can be one measure.

3.2 Preliminary experiments to investigate objective measures

As it is not possible to evaluate all synthetic parts perceptually, an automatic measure is necessary to evaluate the synthetic quality of utterances in order to approximate the quality of each subset solution. However the TTS costs were used as objective measure of signal quality in (Espinosa et al., 2010), it has not been evaluated by a perceptual test. The objective measure should be a good approximation of perceptual evaluation. The correlation coefficient or the ranking correlation coefficient could indicate the reliability of objective measures.

In the following sections, the correlation of objective measures and perceptual qual-

ity will be examined. Afterwards, the use of different ranking measures for spitting greedy will be investigated.

3.2.1 Experimental setup

We proposed some objective measures for quality evaluation of synthetic signals. Some measures such as PESQ (Rix et al., 2001) and Dynamic Time Warping (DTW) between two signals need a reference signal. Basically they evaluate the similarity between a test signal and a reference. Three DTW based measures are proposed: a DTW between Mel Frequency Cepstral Coefficients (MFCC) of the test signal and its natural pair, a DTW between Mel-Generalized Cepstral Coefficients (MGC) of the test signal and its natural pair, and a DTW between MGC features of test signal which is the synthetic signal by using VC and a reference signal which is the synthetic signal by using the full voice corpus. The third DTW calculates the degradation quality of a synthetic test signal from the highest possible synthetic quality using the TTS. The TTS global cost, which is a linear combination of concatenation and target costs, is also proposed to be used as the objective measure of synthetic quality. This cost has also been used in previous works (Chu et al., 2001; Espinosa et al., 2010; Krul et al., 2007; Toda et al., 2006) as the synthetic quality indicator. Moreover it does not need any supplementary computation than the synthesis process in the proposed greedy (the result of the algorithm 1 line 6 can be used directly for line 10).

To investigate the correlation of these objective measures with perceptual quality, a listening test (DMOS) is designed. Six different corpus sizes (75%, 50%, 25%, 10%, 5%, and 1% out of an audio book) are selected randomly as the VC to synthesize the rest of the book.

The listeners are asked to evaluate 60 synthetic samples from each synthetic part. By providing the natural voice of each synthetic signal, the quality degradation of synthetic signal in comparison with the natural voice are asked on a scale from 1 to 5 (5 means without quality degradation and 1 means the lowest quality).

The initial voice corpus contains 3339 utterances of a French expressive audio book (*Albertine disparue* by Marcel Proust) spoken by a male speaker. The overall length of the speech corpus is 10h44. More information on the annotation process can be found in (Boeffard et al., 2012). This audio book will be called *Pod* and used as the voice corpus in the rest of this thesis. The average length of utterances in this corpus

is 120.1 ± 3.2. The average of non-zero f_0 values of voice signal is 90.9 (its standard deviation is 22.9).

3.2.2 Objective measure for synthetic quality

The perceptual test has resulted in 850 evaluation scores. By getting average of annotated scores for each sample, a perceptual score could be assigned to each signal. Two ranking correlation coefficients (Spearman ranking correlation coefficient (Spearman, 1904) and Kendall tau ranking correlation coefficient (Kendall, 1948)) and the Pearson correlation coefficient (Freedman et al., 2007) are calculated between perceptual scores and objective scores. The correlation coefficients between listeners scores and 5 objective measures are compared in table 3.1.

Objective measures	PESQ	DTW-MGC (Natural ref)	DTW-MFCC (Natural ref)	DTW-MGC (Synthetic ref)	TTS global cost
Pearson C.C.	0.07(p>0.2)	-0.41(p<0.001)	-0.38(p<0.001)	-0.40(p<0.001)	**-0.66(p<0.001)**
Spearman R.C.C.	0.08(p>0.1)	-0.39(p<0.001)	-0.39(p<0.001)	-0.40(p<0.001)	**-0.65(p<0.001)**
Kendall tau R.C.C.	0.05(p>0.1)	-0.28(p<0.001)	-0.28(p<0.001)	-0.28(p<0.001)	**-0.48(p<0.001)**

Table 3.1 – The correlation coefficient between objective measures and perceptual evaluation and their p-value.

According to the table 3.1, the TTS global cost has stronger correlation with the perceptual score than PESQ or DTW on different acoustic features (MFCC, MGC).

While the reported correlation coefficients are calculated on synthetic signals with 6 voice corpus sizes, the mean of perceptual and objective scores on each voice corpus size could reveal more information. The impact of corpus size on synthetic quality (with perceptual and objective measures) is investigated. Figure 3.3 compares the objective and perceptual score for synthetic utterances in different corpus size (out of a 10h44min voice). The horizontal axis indicates the size of VC which is selected randomly.

The increasing trend of DMOS score and decreasing trend of TTS global cost for larger VC confirm that the quality of synthetic signals with larger voice corpora will be improved. Nevertheless, the perceptual quality of synthetic signals with 25%, 50%, and 75% of the full corpus (more than 1 hour) are not significantly different (three right red bars). It means that using more data for TTS voice corpus after a threshold will not improve the signals quality enough to be distinguished by human perception.

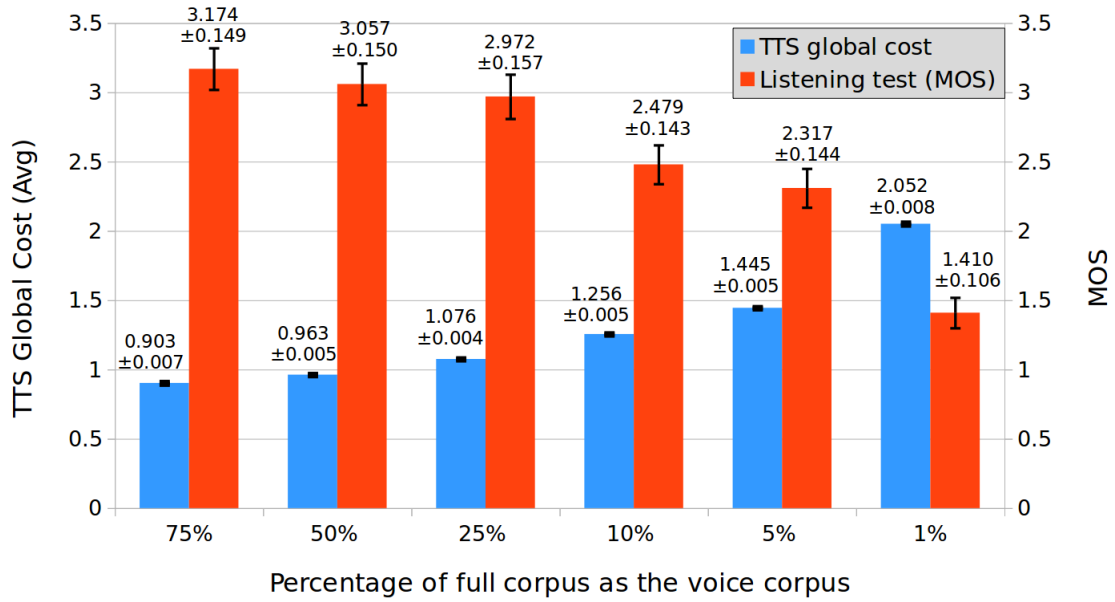


Figure 3.3 – The TTS global cost and perceptual score for different voice corpus sizes (subcorpora are extracted randomly).

Although it is observed only on this TTS and for random corpus reduction. For a general conclusion this hypothesis should be tested on other TTS systems as well.

These results show that the TTS global cost could be used as the approximation of perceptual quality.

3.2.3 Ranking measure

Beside the objective measure of quality, an objective measure for ranking candidates is needed. The ranking measure is used to decide which utterance should be removed from VC at each step. Like in algorithm 1, these two measures can be the same but we propose some other measures for ranking candidates. The DTW and PESQ are computationally expansive so they do not seem to be practical for the ranking. Ranking measures can be acoustically based such as the TTS global cost or the usage frequency of utterance's diphones (based on (Krul et al., 2007)) to synthesize the rest of the book. Some other measures like diphone entropy (Nose et al., 2015b), or diphone Kullback–Leibler divergence (KLD) (Krul et al., 2006) can be also used for ranking utterances. These proposed objective measures are listed in table 3.2.

By following the proposed spitting greedy on a small corpus (334 utterance) the

Ranking measures	Description
TTSCost	The TTS global cost (normalized by utterance length) which would result from the synthesis process
diphUsage	The number of times that the utterance's diphones are used by TTS to synthesize the synthetic part (normalized by utterance length)
EntropyS	The entropy of diphones in SP after adding the candidate
EntropyV	The entropy of diphones in VC after removing the candidate
KLD(S‖V)	The KLD between diphones distribution in SP plus the candidate and VC without the candidate
KLD(V‖G)	The KLD between diphones distribution in voice corpus without the candidate and the full corpus
KLD(S‖G)	The KLD between diphones distribution in SP plus the candidate and the full corpus

Table 3.2 – The different measures for ranking candidates.

performance of these measures are evaluated. The TTS global cost of synthetic part (rest of the corpus) is taken into account as an approximation of synthetic quality. Although reducing the size of voice corpus would result in higher TTS global cost, the rate of increasing TTS global cost indicates the performance of the ranking measures.

In order to consider the impact of ranking measures, we rank candidates based on maximization and minimization of these measures. For example, by ranking utterances based on minimum TTS global cost, as the ranking measure, the TTS global cost of synthetic part, as the quality measure, would be minimum. In an opposite way, by ranking utterances based on maximum TTS global cost, the reduction process would result in a solution with lowest synthetic quality. It is expected that the random selection method achieves a solution whose synthetic quality is between the highest and the lowest possible synthetic quality. They are respectively obtained by the minimum and maximum TTS global costs as objective measure. This gap between maximization and minimization of ranking measures, which surrounds the random solution, could reveal their performance.

The figure 3.4 shows the sum of synthetic part TTS global costs when greedy algorithm employs different ranking measures for reduction of a voice corpus with 334 utterances.

A lower line in the figure 3.4, shows that TTS global cost of synthetic part is lower

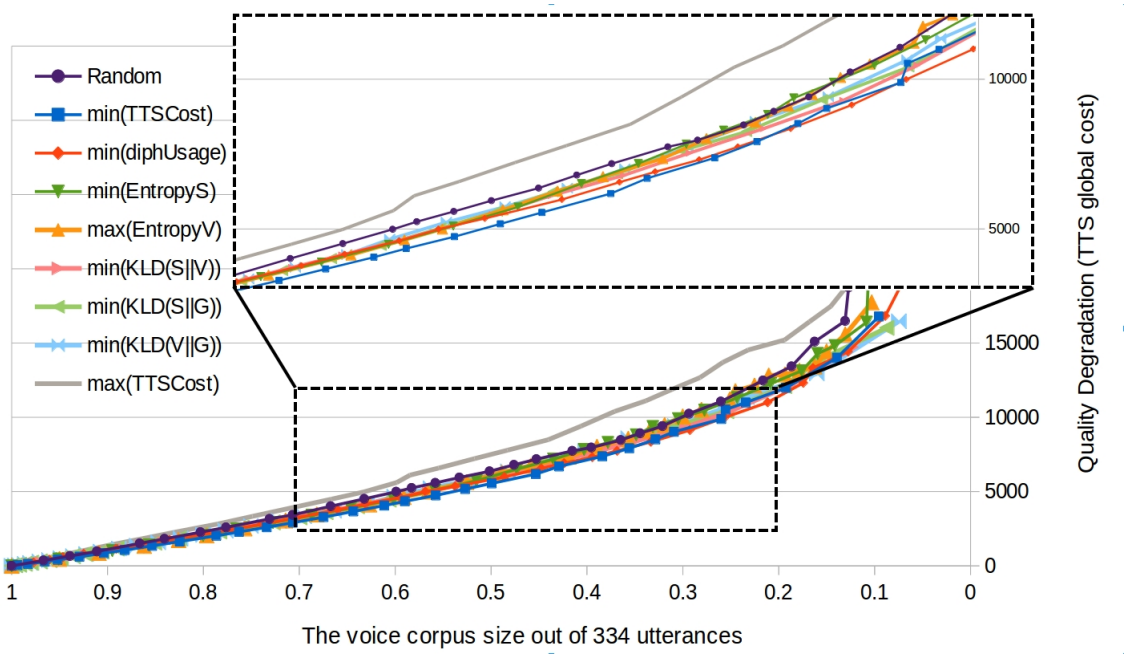


Figure 3.4 – The performance of different ranking measures based on the quality degradation in the spitting greedy.

in different voice corpus sizes. So the lowest line would be interesting because by reducing voice corpus size, the quality degradation would be the lowest. The result did not reveal any advantage of other ranking measures compared to the TTS global cost since they are closer to the random line. Although these results are achieved on a small voice corpus (1 hour), the TTS global cost achieved lower quality degradation in voice reduction process than the random strategy. Moreover by selecting this measure some computations would be gained as the TTS global cost value is already computed in the algorithm 1 line 6 and 10.

3.3 Evaluation of spitting greedy

In this section the evaluation results of proposed corpus reduction method will be detailed. Based on the previous perceptual test results, the objective measure for ranking utterances and evaluation of synthetic quality in the proposed spitting greedy is the TTS global cost.

Although the main problem in our case is to synthesize the rest of the book, a fixed synthetic part as *test section* would help to compare different methodologies for corpus

design. We assume that since the test part came from the same book, the synthetic quality of this part can be generalized to the rest of the script. The initial corpus is the same audio book as what has been described in the previous section (see the end of section 3.1). The audio book has been divided into two parts. The test section $\mathcal{T}$ which is randomly selected as a continuous part with 334 utterances (10% of the whole audio book). The rest of the audio book is named the full corpus and is denoted $\mathcal{F}$ in the remainder. The TTS can use $\mathcal{F}$ or just a certain percentage of $\mathcal{F}$ as VC to synthesize $\mathcal{T}$. The voice corpus reduction is done based on the spitting greedy and a random strategy. Table 3.3 shows the corpus reduction rates used for synthesis.

The size VC out of $\mathcal{F}$	100%	70%	40%	15%	7%	3%
Num of diphones in VC	362126	253488	144850	54318	25348	10863
Num of utt. in VC (greedy)	3005	1941	1137	478	228	63
Num of utt. in VC (random)	3005	2098	1194	435	186	57

Table 3.3 – The reduction rates and the length of VC in terms of number of diphones and utterances. The VCs result from the spitting greedy and random methods.

Two perceptual test are designed to evaluate the quality of signals which are synthesized using voice corpora obtained by the proposed spitting greedy. The first perceptual test is designed to investigate the impact of voice corpus reduction by spitting greedy on synthesizing quality. The purpose of the second perceptual test is to compare the performance of spitting greedy to random voice corpus reduction in terms of synthetic quality. In the following, those two tests are presented.

3.3.1 Impact of voice corpus reduction on synthetic quality

Based on the voice corpus reduction rates in table 3.3, $\mathcal{T}$ has been synthesized. Since the IRISA TTS system is unit selection-based, some utterances may failed to be synthesised, specially in small voice corpus size. After removing these uncommon samples 70 utterances have been selected randomly. In order to have perceptual samples with an acceptable duration, some utterances have been concatenated or cut. If the length of selected synthetic signal is less than 4 seconds, the next utterance in text order is concatenated. Then, the synthetic signals are split in first 6 seconds to be used as listening samples. Samples from 6 voice corpus sizes and two corpus reduction methods are used to design a MUSHRA test (Recommendation, 2003). For each

step, the overall quality of 11 synthetic signals have been asked to evaluate on a scale from 1 to 10 with a step of 1. Each sample in a test step corresponds is obtained with different size and method. Synthetic signals and corresponding natural voice, which have same script, are available to listeners. The listeners are asked to do 10 steps of MUSHRA test after an introduction step. The estimated time for doing this test is 25 minutes.

This perceptual test is done by 14 listeners which provides 1441 evaluation scores for synthetic signals. To investigate the impact of corpus size on synthetic signals, the average score for each size/method has been calculated. The figure 3.5 shows the results of the proposed test. This figure demonstrates that the average score for the different voice corpora are in the same level. It indicates that not only the quality of synthetic signals based on random and greedy strategy are not significantly different, but also reducing the voice corpus size has no significant impact on the output quality, at least until a reduction down to 15% of the full corpus.

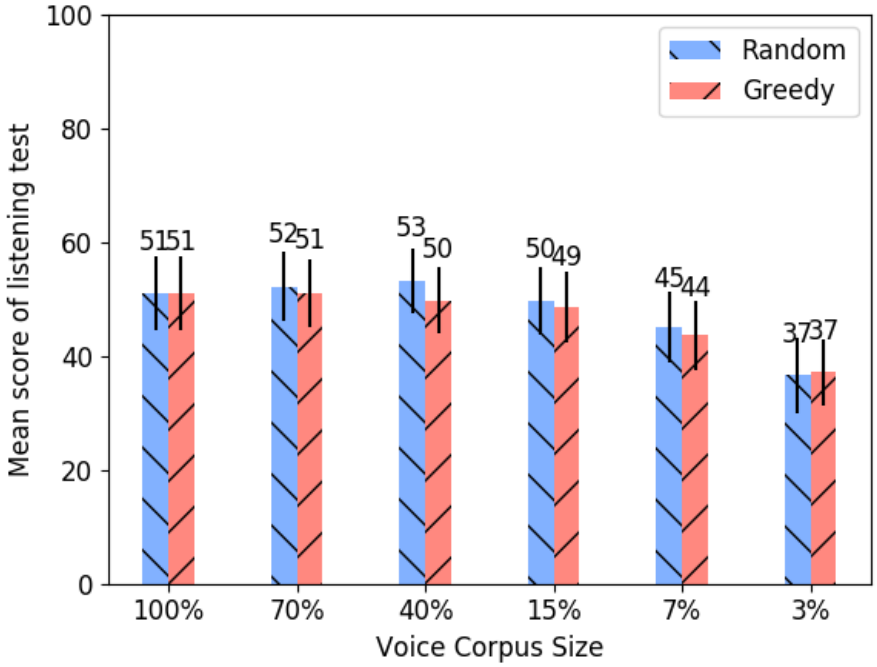


Figure 3.5 – Mean opinion score for the proposed greedy and random methods considering different voice corpus sizes.

According to listeners feedback, we found out that comparing 11 samples is a difficult and exhausting task. This problem encourages us to estimate the preference of listeners as if they were asked to compare two signals. So the resulting scores from

MUSHRA test are used to simulate an AB test. Concretely, each two signals are compared based on their perceptual score. The score values of two signals are converted to a simple comparison in order to simulate the preference of listeners and if the scores are equal it is assumed the listeners have not preference. Results of this conversion are displayed in the figure 3.6. The numbers in the heat-map table indicate the preference percentage of vertical labels in comparison to horizontal labels.

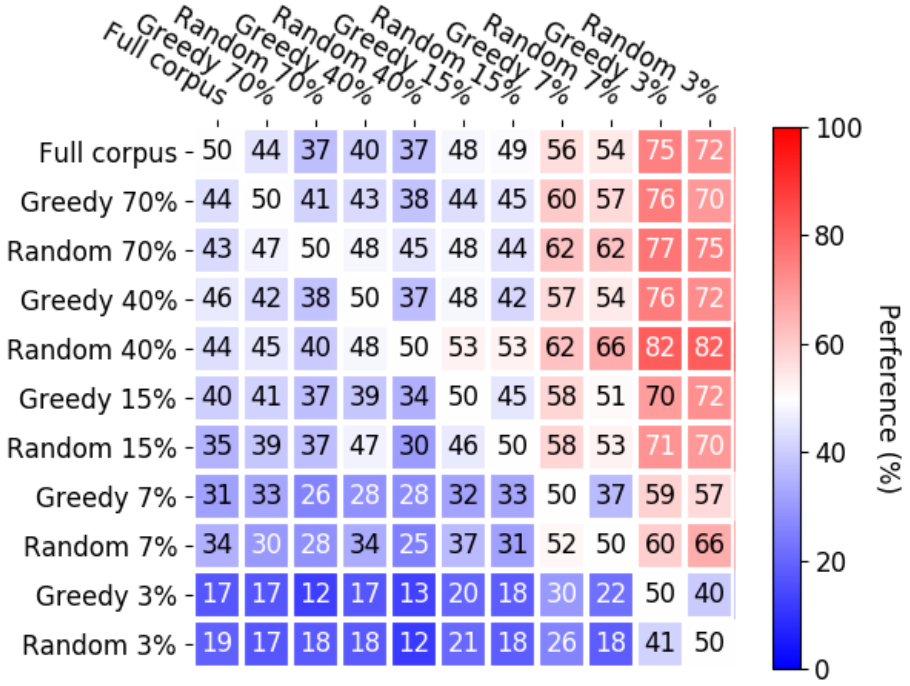


Figure 3.6 – Listeners preference obtained from the MUSHRA test to compare modified spitting greedy and random method for voice corpus reduction.

Based on this figure, the preference of synthetic signals with small voice corpus (left-down) is lower than synthetic signals with large voice corpus (right-up). It confirms that voice corpus reduction decreases the TTS synthetic quality. By looking at cells in large corpus sizes (left-up), it can be observed that the preference numbers for corpus size bigger than 15% are around 50. This observation confirms the hypothesis in section 3.2.2.It means after a certain voice corpus size the quality of synthetic signal is not improved perceptually by increasing the voice corpus size.

Both figures 3.5 and 3.6 do not show superiority of spitting greedy in comparison with random strategy. This is contradictory with what we expected based on previous studies such as (Chevelu & Lolive, 2015). As the MUSHRA test has been reported to

be a difficult task for this comparison, another perceptual test is proposed for comparing the performance of these two corpus reduction methods.

3.3.2 Performance of spitting greedy vs. random selection

Based on listeners' feedback from previous perceptual test, some modifications have been done on samples preparation and the test platform. While we use same test section $\mathcal{T}$ and reduction rates (table 3.3), the final listening signals are prepared in a different way. The utterances have been synthesized from the beginning until the first speech pause after 90 diphones. In this way, all samples for sizes/methods will have same content. The duration of samples are between 5 to 10 seconds. Among 334 utterances of test section, 70 samples have been selected for the listening test according to the highest acoustic distance (Chevelu, Lolive, et al., 2015). The acoustic distance is computed by DTW on MGC features of the two signals. As reported in (Chevelu, Lolive, et al., 2015), this selection method helps to focus on the most different samples.

An AB test has been prepared with 40 steps. For each step, listeners are asked to give their preference in terms of overall quality between two synthetic signals. These signals have been synthesized using different voice corpora but with same size. Voice corpora are a sub part of $\mathcal{F}$ obtained from the random strategy or the proposed spitting greedy. The estimated time for doing the whole test is 15 minutes.

The listening test has been done by 9 listeners. For each voice corpus size between 66-70 comparisons have been achieved. Out of 340 comparisons in total, the random strategy has been preferred 132 times, the greedy strategy has been preferred 118 times, and 90 times listeners selected no preference. The figure 3.7 shows the percentage of preference for corpus reduction methods for different voice corpus sizes.

The figure 3.7 does not reveal any significant superiority of the modified greedy. Even the synthetic signals for 15% of full voice corpus with random strategy have been evaluated slightly better than the modified greedy.

In order to investigate the impact of selecting $\mathcal{T}$ for perceptual test, the TTS global cost of the AB test's samples is displayed in the figure 3.8.

The TTS global cost of listening test signals given by the random selection are not significantly different from those given by the proposed greedy on the test section. The same trend is observed for the rest of the book (synthetic part). Given those results,

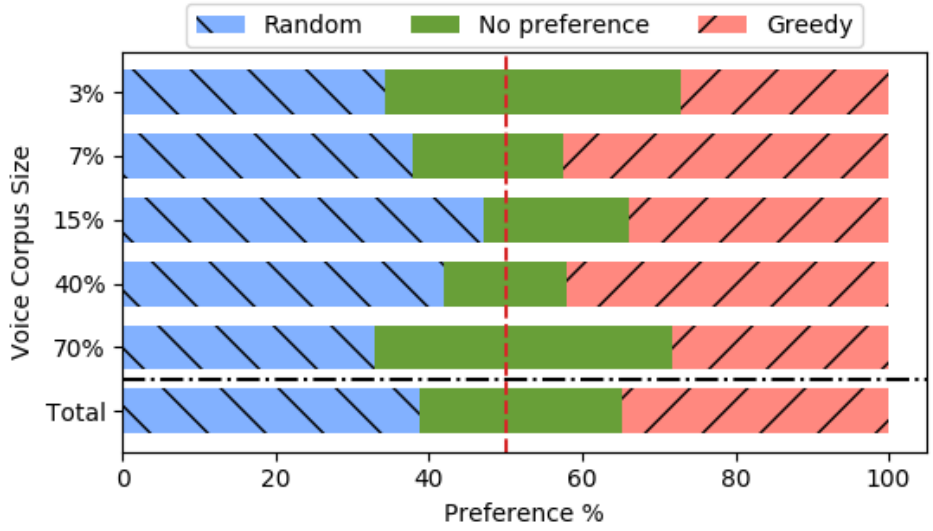


Figure 3.7 – AB test results between random strategy and greedy strategy for reducing voice corpus. The y axis indicates the ratio of VC out of $\mathcal{F}$.

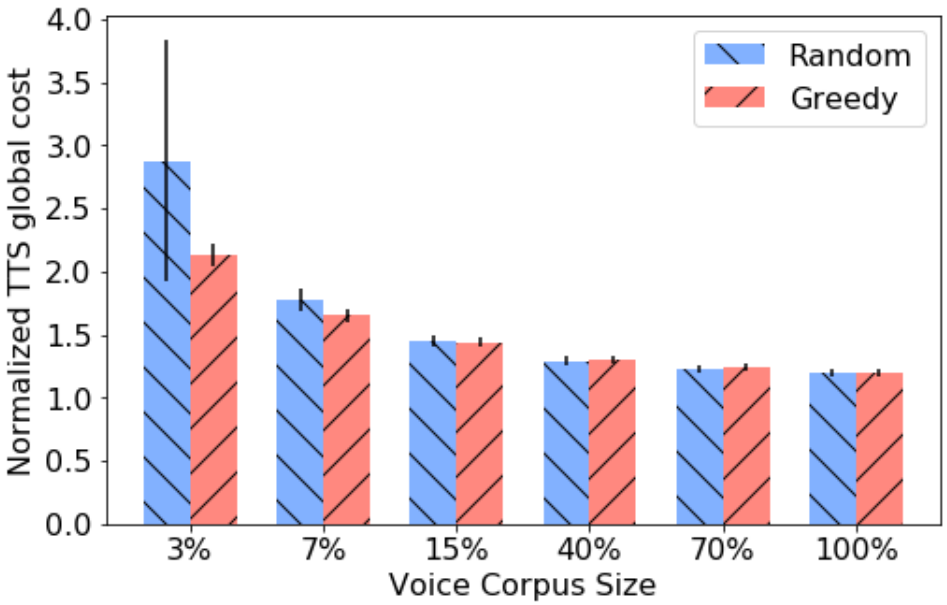


Figure 3.8 – The normalized TTS global cost of listening test samples.

we can conclude that the random reduction works as well as the proposed spitting greedy. The explanation could be the approximation level of proposed method (level 3 in table 3.2). It means that reducing the computational time costs a lot in terms of the optimality of solution. Hence the performance of subset solution becomes close to a

random selection.

3.4 Conclusion

In this chapter, a posterior strategy has been followed. A modified spitting greedy algorithm has been proposed to reduce a given voice corpus. The result of voice reduction process could lead to find the linguistic characteristic for script selection problem in voice corpus design.

The computational time has been the main challenge in this subset selection problem. By modifying the original spitting greedy, the computation of the algorithm has been reduced to a reasonable time. However this approximation level costs lower efficiency and makes the solution closer to a random selection.

In the first step, some objective measures like PESQ, DTW between synthetic signal and voice signal, and TTS global cost have been investigated. A perceptual listening test showed a higher correlation between TTS global cost and perceptual quality. Afterwards, the TTS global cost has been compared with some other linguistic metrics for ranking measure. By running greedy algorithm with these ranking measures on a small voice corpus with 334 utterances, no significant superiority of these linguistic measures have been observed. Therefore the TTS global cost has been employed in greedy algorithm for ranking candidates in each reduction step.

In a MUSHRA test, the random strategy and proposed greedy are compared for different voice corpus sizes. It has been observed that after a certain size of voice (1 hours of our audio book), the voice corpus is big enough and the difference of synthetic signals can not be distinguished perceptually. Moreover any differences between random and proposed greedy has been observed. In order to evaluate the performance of proposed greedy another AB preference test has been run. The result of this listening test confirmed that listeners did not prefer the signals which are synthesized using voice corpus obtained with the proposed greedy in comparison with a random strategy.

To sum up this chapter, we did not find an algorithm, which has a reasonable computational time and performs better than random, to follow the posterior strategy. Despite that, the TTS global cost has been found to be a good measure to approximate the synthetic quality.

In the next chapter, the script selection problem will be investigated based on linguistic information.

CHAPTER 4

PHONEME-EMBEDDING BASED APPROACH

After the unsuccessful posterior strategy in the previous chapter, we propose to design the voice corpus by following script selection strategies. The goal is to select a subset of book script based on linguistic information.

The main idea of this chapter is to derive a vector representation of the linguistic information in order to facilitate the selection of a subset of utterances having a good linguistic variety from a text corpus. Increasing the number of features and samples leads to an exponential growth of the covering size if no feature selection is done. Instead of introducing expert knowledge to select the features, we propose to use a model for that task. Deep neural networks and particularly deep auto-encoders could be used to do so. In our case, we propose a Convolutional Neural Network (CNN) (Lecun et al., 1995) to map utterances to an embedding space. Then, we try to find a tiling of the embedding space, in order to obtain the largest possible linguistic covering, that could improve the speech synthesis quality compared to standard approaches. These selection approaches are compared to LSTM (Sutskever et al., 2014) and Doc2vec (Le et al., 2014) methods as well as to a standard set covering one, implemented as the covering of all diphonemes using a greedy strategy (Barbot et al., 2015; Chevelu & Lolive, 2015). The perceptual evaluation shows that the proposed methods are more efficient than the standard one. Moreover, a crucial asset of these embedding-based approaches is that it is not necessary to select features, they adapt automatically to the book to be generated.

This experiment has been published as conference papers in (Shamsi et al., 2019a, 2019c).

4.1 Embedding-based corpus design

The proposed approach relies on a CNN with the aim of learning a non-linear transformation from textual and linguistic data into a new pertinent representation without manual feature extraction/selection. The derived utterance embedding enables to guide and compare some selection algorithms to extract a set of utterances as a subset offering a large linguistic richness.

Figure 4.1 shows the process of corpus design: (1) information extraction from the text corpus, (2) projection of feature vectors into an embedding space, and (3) utterance selection.

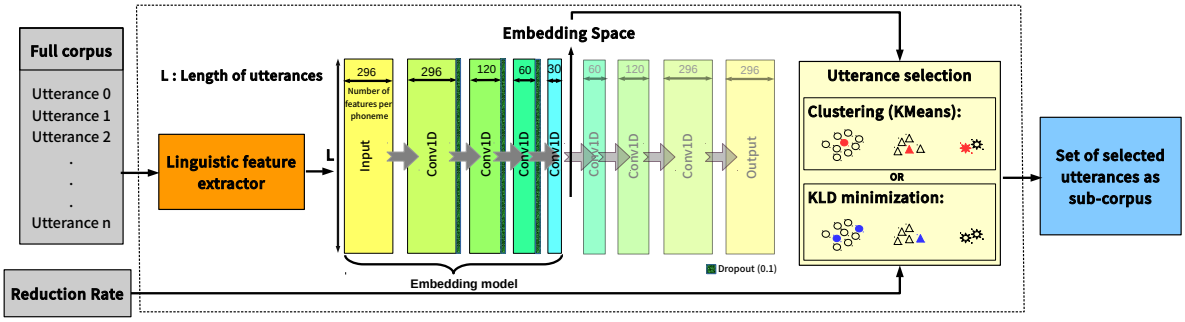


Figure 4.1 – Corpus design process and CNN auto-encoder architecture.

4.2 Information extraction

We define a linguistic feature vector, for each phoneme in the text utterance, providing information about the phoneme, e.g., its identity, preceding and following neighbours, its position in the syllable/word/utterance it belongs to, etc. The linguistic features are automatically extracted (Perquin et al., 2018) from the corpus. Thus, the linguistic vector, of size 296, contains categorical and numerical features. The categorical attributes represent information about quinphonemes, syllables, articulatory features, and POS for the current, previous and following words. These features are converted to a one-hot vector. The numerical features take into account information such as the phoneme position inside the word or utterance. These numerical features are normalized so that all the entries of the linguistic vector are in the range $[0, 1]$. The linguistic content of an utterance is then represented by the sequence of linguistic

feature vectors associated to the phonemes that compose it.

The proposed linguistic features (*Ling inf*) has been compared with only diphones identity (*Diph tag*) as a one-hot vector. The performance of embedding models using these different feature types shows using more information (*Ling inf*) could be helpful.

4.3 Embedding model

From this initial representation of the linguistic content at phoneme and utterance levels, using an embedding space enables to derive a continuous and compressed representation. Importantly, this approach avoids the injection of expert knowledge to drive the selection of the most important features, letting the model reveal what is of interest.

To build up this embedding space, an auto-encoder based on a multi-layer CNN has been implemented, as shown on Figure 4.1. To avoid overfitting, a dropout layer is used with a 0.1 drop probability after each layer in the encoder (Srivastava et al., 2014). CNN layers are used with kernel size of 5 and the $\tanh$ activation function. The loss function is the Mean Squared Error (MSE).

4.3.1 Training sample types

Three types of sample sets have been tested to train the CNN auto-encoder: a set of utterances (*Utt*) with variable length, a set of chunks provided by a sliding window (*SlidWin*) of size 100 phoneme instances with a step size of 10 phonemes, and a set of breath groups (*BG*) with variable length. The length of *SlidWin* samples is around the average length of utterances in corpus. Consequently, after training with *SlidWin* samples, the *Utt* samples can be used for prediction and to transform its phonemes' features to embedding features. It helps to stay at the utterance level to compute the embedding vectors.

Table 4.1 shows the number of samples and their average length (number of phoneme instances) which are used for training with the different sample types.

After training, the network is used to generate, for each input sequence of linguistic vectors at utterance level, a sequence of unit vectors in embedding space. Its length is equal to the number of phoneme instances in the input utterance (or breath group).

Sample type	Sample number	Avg. length (in phonemes.)
Utt	3005	120.5
BG	10287	35.2
SlidWin	36203	100

Table 4.1 – Number and average length of samples

4.3.2 Other embedding architectures

To compare our proposed embedding model to state of the art models, 3 other architectures have also been employed. An fully connected multilayer perceptron (*MLP*) as a phoneme auto-encoder with a bottleneck, a LSTM model based on (Sutskever et al., 2014) and *Doc2vec* are implemented.

The *MLP* model uses only the phonemes information as the input and the output, while the other models get a sequence of phonemes corresponding to utterances, breath groups, or *SlidWin* samples. The *MLP* auto-encoder does not profit from contextual information, otherwise the training process would be expensive. Due to this lack of information, the performance of *MLP* auto-encoder is not as good as other methods.

The long short-term memories (*LSTM*) based model (Sutskever et al., 2014), which is a sequence-to-sequence model, has one LSTM layer of encoding and one LSTM layer of decoding. However this proposition was used for translation task, the LSTM hidden states can be used as embedding vectors for utterances when the model has been trained as an auto-encoder. The idea is similar to the one in (Mueller et al., 2016) which used a LSTM based model for semantic similarity of utterance.

The *Doc2vec* model (Le et al., 2014) is learnt using the *gensim* toolbox with a window size of 5 and a minimum count of input vectors equal to two.

4.4 Selection Method

The main idea behind utterance selection is to extract a set of utterances from a book that offers a representative linguistic coverage while limiting the linguistic unit repetitions. In our case, the term unit stands for phonemes in context, based on the linguistic features used. The concrete goal is to provide a large variety of options to the TTS system while minimizing the voice size. We propose three methods for selecting utterances: the two first methods are based on a clustering approach, the third one

tends to reach a target distribution of linguistic events.

4.4.1 Phonemes clustering followed by set covering

Not only the unit identity could carry important information for voice corpus but also the contextual information. The unit definition is extended with other linguistic information like the features which have been mentioned in section 4.2. Phonemes in embedding space contain all linguistic information in the form of vectors of continuous values. A representation of phoneme instances in embedding space is displayed in figure 4.2.

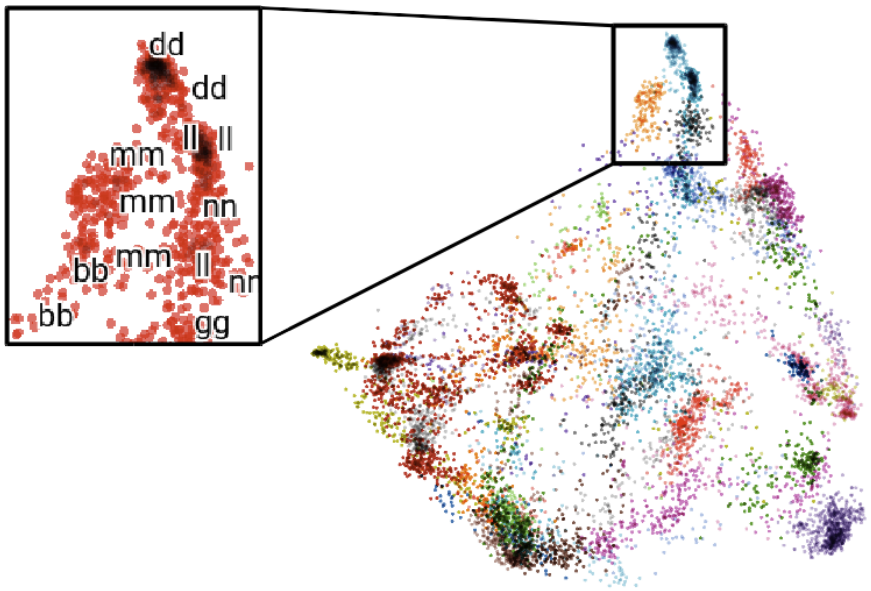


Figure 4.2 – Phoneme instances projected in a two dimensions embedding space. Each color corresponds to phoneme identity.

The idea behind this method is to cover the variation of phonemes' embedding representation. Firstly, the new representation of phonemes will be clustered into new categories. These new categories of phonemes take into account the contextual information. The K-Means algorithm is employed for this clustering. Afterward as the second step, a coverage method is expected to find a subset of categories' representation. A greedy process is used to extract the best subset for coverage of new phonemes' category. The rest of coverage method to have an exact length of subcorpus is same as classic greedy set covering which will be explained in section 4.5.2.

The number of clusters could be controlled by the limited length of sub-corpus, which is set to 2000 (the best of 1000, 2000, 4000, 8000) in our problem. This method will be called *KMeansSC*.

4.4.2 Utterance clustering

Clustering the utterances could categorize utterances based on their similarity. By selecting one utterance per cluster, we assume that it represents the information of other utterances of its cluster. In particular, one may consider that the most representative utterance is the closest one to the cluster center.

In order to compute a similarity measure between utterances with different lengths, we have built a numerical and fixed dimensional representation of utterances. Let us consider an utterance u composed of m phoneme instances, its i^{th} phoneme instance is represented by the embedding vector $p_i = (x_1^i, \ldots, x_N^i)$, where N corresponds to the embedding dimension. Several aggregation operators could be used to take into account the contributions of phonemes in u, like the sum or average. We have chosen to use the average to avoid the utterance length-dependency: u is then represented by $\hat{u} = (f_1, \ldots, f_N)$ where $f_j = 1/m \sum_{i=1}^{m} x_j^i$.

The clustering of the full text corpus $\mathcal{F}$ is made based on the K-Means algorithm using the Euclidean distance between utterance vectors $\hat{u}$ as the similarity measure (The cosine distance gave similar results). As mentioned above, the closest vector to the cluster center is selected from each cluster. The length l_{VC} of the set VC of selected sentences (as the voice corpus) is given by the sum of the length of its elements (in terms of number of phoneme instances). In order to achieve a target reduction rate τ^* of $\mathcal{F}$, the cluster number is iteratively updated (the selection is then redone): its initial value K_0 is set to $\lfloor \tau^* \times$ (number of utterances in $\mathcal{F}) \rfloor$; resulting from step i, a selected subset VC_i is derived using K_i clusters and K_{i+1} is set to the $\lfloor K_i \times \tau^* \times l_{\mathcal{F}} / l_{\mathcal{S}_i} \rfloor$. This selection method will be referred by *KMeans* in the remainder.

4.4.3 KLD minimization

A greedy strategy to minimize the Kullback-Leibler divergence in the context of corpus design has been proposed in (Krul et al., 2006). Although this method was based on the phonological unit distributions, the idea can be transposed to continuous values

in embedding space. In our case, the target distribution is given by the unit distribution in the full corpus $\mathcal{F}$ or test section $\mathcal{T}$.

Precisely, for each dimension of the embedded phoneme vectors, values are normalized to the range $[0, 1]$ and an histogram h is then computed by binning the values into ten bins ($X = \{[0, 0.1), \ldots, [0.9, 1]\}$). Thus, for each latent feature f_j, its probability distribution can be defined using the associated histogram $h(f_j)$. The KLD between the probability distribution P_s^j of f_j in the selected set of utterances VC and the probability distribution P_t^j in the target set of utterances is derived as follows:

$$KLD(P_{\mathcal{S}}^j \,||\, P_t^j) \;=\; -\sum_{x \in X} P_s^j(x) \log\left(\frac{P_s^j(x)}{P_t^j(x)}\right).$$

To achieve a target sub-corpus size, at each iteration, a greedy process selects the utterance which minimizes the average of KLDs (one KLD per feature) between the target distribution and the distribution computed from the new set of utterances, including the candidate utterance. This selection method will be named *KLD*.

4.5 Experiments and results

The original audio-book generation problem goal was to synthesize the rest of the book (see section 2.1). As it has been mentioned in section 3.3 a *test section* helps to compare different voice subcorpora results.

In this section, first the experimental setup to evaluate script selection methods will be described. Afterwards, these methods will be compared based on TTS costs of *test section* synthetic signals. Finally the result of perceptual comparison between best configurations will be brought up.

4.5.1 Experimental setup

The initial corpus is *Pod* corpus which has been introduced in 3.2.1. The audio book has been divided into two parts. A test section $\mathcal{T}$ which is randomly selected as a continuous part with 334 utterances (10% of the whole corpus). The rest of the audio book is named the full corpus and is denoted $\mathcal{F}$ in the remainder. $\mathcal{F}$ is composed of 3005 utterances and 362126 phoneme instances. The objective is to extract from $\mathcal{F}$ a subset VC of a given size. The natural signal samples of VC will be used to

synthesize the utterances of $\mathcal{T}$ by the IRISA TTS system (Alain et al., 2017). To derive the embedded representation of utterances of $\mathcal{F}$, 90% of $\mathcal{F}$ are used for training the CNN models and 10% are used as a validation set to avoid overfitting.

4.5.2 Best configuration selection

Several embedding sizes have been tested ($N = 240, 120, 60, 30, 15$). Table 4.2 display the reconstruction error of CNN auto-encoder models. We can observe that $N = 30$ gives the best reconstruction error for the CNN models.

Embedding size	SlidWin_Utt *	Utt_Utt	BG_BG
15	0.00035	0.00077	0.0093
30	**0.00021**	**0.00067**	**0.0091**
60	0.00014	0.00066	0.0106
120	0.00021	0.00072	0.0121
240	0.00019	0.00105	0.0135

Table 4.2 – Reconstruction error (MSE) of CNN auto-encoder with different embedding sizes and training sample types. (* The first sample type is for training and the second is for prediction.)

In order to compare the performance of the selection methods and evaluate the impact of the selection size on the synthesised speech quality, several sub-corpus sizes of $\mathcal{F}$ have been tested: 50%, 40%, 30%, 20%, and 10%. Based on previous experiments, as synthetic signals given by a large voice corpus can not be distinguished perceptually, we have avoided to evaluate sub-corpus sizes above 50%. Consequently, the selection methods under comparison are the following:

— *Random*: the baseline method is a random selection of utterances. To have representative results, 10 random selections have been built for each reduction size, and for the evaluation, the average values are considered.

— *SC*: this system is based on a greedy strategy to solve a Set Covering problem (Barbot et al., 2015). The utterances are selected so as the solution under construction covers at least η times each linguistic feature. Starting from 1, η is incremented until the target sub-corpus size is reached.

— *GreedyKLD*: a greedy algorithm is used to minimize the KLD between the diphoneme distribution of the selected subset VC and a target distribution as

done in (Krul et al., 2006). The target distribution can be diphoneme distribution in $\mathcal{F}$ (*KLD(Full)*) or in $\mathcal{T}$ (*KLD(Test)*).

— *Doc2Vec/LSTM/CNN_KMeans*: as detailed in sections 4.3 and 4.4.2, the selection strategy is the K-Means algorithm which clusters the embedding space. This embedding is derived by Doc2Vec model or LSTM auto-encoder, which are presented in Section 4.3.2, or a CNN auto-encoder.

— *CNN_KMeansSC*: as it is described in section 4.4.1, the K-Means algorithm clusters the phonemes in the embedding space, which is given by a CNN auto-encoder. the same algorithm as *SC* is applied, but instead of using linguistic features, new cluster labels has to be covered.

— *CNN_KLD*: it is a variant of *GreedyKLD*. The considered distributions are those associated to the embedded vectorial representation as explained in Section 4.4.3.

All methods which have been tested for the voice corpus reduction in this experiment are listed in table 4.3.

	Selection method	Embedding	Embedding size	Type of Information	Training Samples
Baseline	Random	-	-	-	-
	SC	-	-	Diph tag, **Ling inf**	-
	GreedyKLD	-	-	Diph tag	-
Embedding based	KMeansSC	CNN	15,30,60,120,240	Diph tag, Ling inf	Utt, BG, SlidWin
	KMeans	Doc2Vec	15,30,60,120,240	Diph tag	Utt, BG, SlidWin
		LSTM	15,30,60,120,240	Diph tag, Ling inf	Utt, BG, SlidWin
		MLP	15,30,60,120,240	Diph tag, Ling inf	Utt, BG, SlidWin
		CNN	15,**30**,60,120,240	Diph tag, **Ling inf**	**Utt**, BG, SlidWin
	KLD	MLP	15,30,60,120,240	Diph tag, Ling inf	Utt, BG, SlidWin
		CNN	15,**30**,60,120,240	Diph tag, **Ling inf**	Utt, BG, **SlidWin**

Table 4.3 – Corpus design methods with different configurations.

In the remainder, embedding based methods for script selection are named as follows; (embedding model)_(training sample type)_(selection method).

For each selection method and reduction size, the obtained voice is used to synthesize the utterances of $\mathcal{T}$. Figure 4.3 displays the associated average TTS global cost (the average concatenation and target costs are not detailed here since they indicate the same trends). According to the previous experiment, the TTS global cost is used as an approximation of perceptual quality. We can observe that the reduced set provided by *CNN_Utt_KMeans* achieves the best performance. We have also compared the proposed CNN embedding to the Doc2Vec and LSTM models. The results show, on TTS global cost, that the CNN based approach performs better.

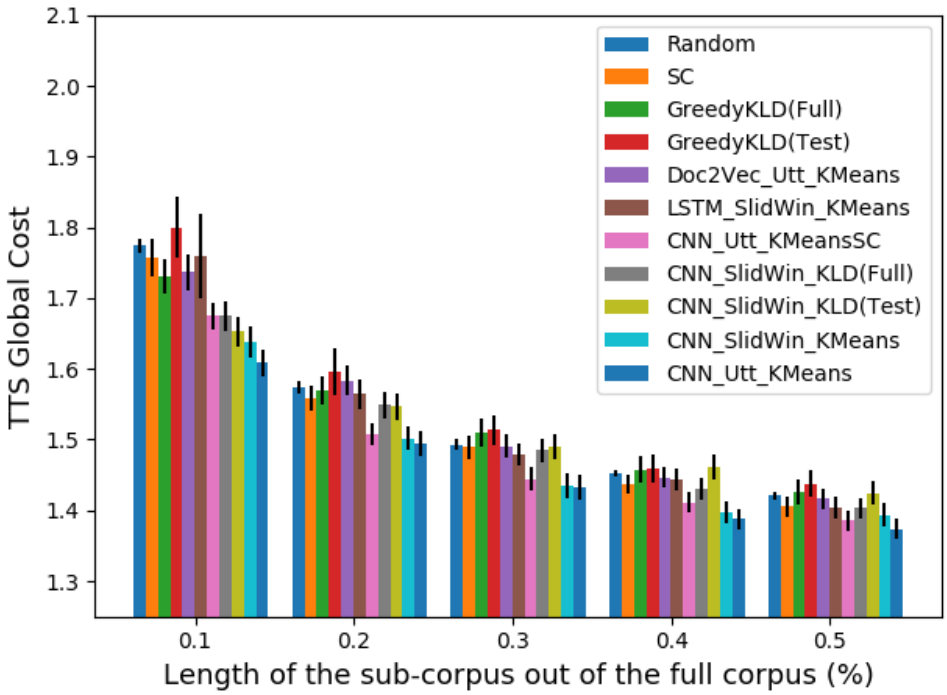


Figure 4.3 – TTS global cost of synthetic signals using different voice subcorpus resulted by best configurations/systems.

Considering TTS global cost results, we keep the two following approaches, relying on different selection strategies, for further evaluations: *CNN_Utt_KMeans* as the coverage based method and *CNN_SlidWin_KLD(Full)* as the distribution based method.

4.5.3 Subjective evaluation

Based on objective measures, three methods have been chosen to be compared perceptually: *SC*, *CNN_Utt_KMeans* and *CNN_SlidWin_KLD(Full)*. The *SC* is selected as the best in the state of the art methods. The *CNN_Utt_KMeans* and the *CNN_SlidWin_KLD(Full)* are selected as the two of best utterance selection methods. The utterances of the test section have been synthesized using 10% of $\mathcal{F}$ selected by each of these methods for perceptual test. Three AB preference tests have been conducted to compare the following pairs of systems:

1. *CNN_SlidWin_KLD(Full)* and *SC*, 19 listeners
2. *CNN_Utt_KMeans* and *SC*, 17 listeners
3. *CNN_Utt_KMeans* and *CNN_SlidWin_KLD(Full)*, 13 listeners

Each test is composed of the 100 samples with the highest DTW on MCep features from the test set (Chevelu, Lolive, et al., 2015). The samples are shorter than 7 seconds. Listeners were asked to compare 30 pairs in terms of overall quality. The results are reported on Figure 4.4.

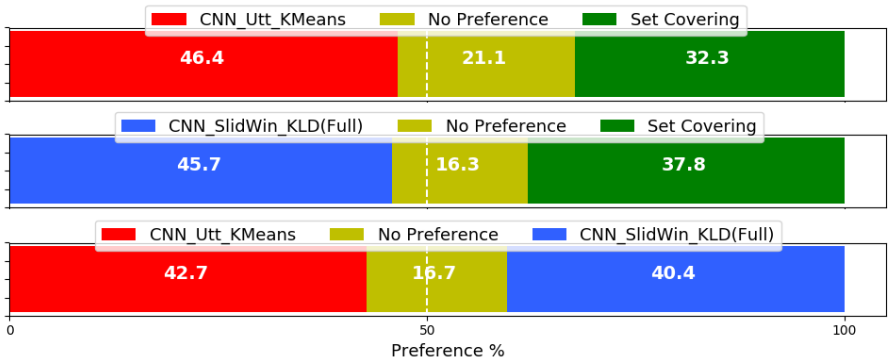


Figure 4.4 – Listening test results of comparisons between synthetic signals achieved by different text selection methods for TTS voice corpus.

Synthetic signals provided by *CNN_Utt_KMeans* and *CNN_SlidWin_KLD(Full)* are judged to be of better quality than the ones from *SC*, which confirms the ranking between these methods provided by the objective measures. Moreover, listeners have a small preference for the *CNN_Utt_KMeans* method rather than the *CNN_SlidWin_KLD(Full)* one but this trend is not really significant. These results indicate that the CNN auto-encoder as the feature selection/extraction method is at least as efficient as state of the art methods while it does not need manual feature selection.

4.6 Conclusion

In this chapter, we have presented a method for voice corpus selection. In the framework of TTS corpus design, we have showed that a CNN auto-encoder can be used successfully to extract linguistic information. The K-Means clustering and the KLD methods work properly using embedded representations achieving better results than random, or even than the best state-of-the-art methods such as greedy based set covering algorithm. We have also compared the proposed CNN embedding approach to *LSTM* and *Doc2Vec*, and it proves to work better in the particular context of corpus design. The subjective evaluation has confirmed this result showing a preference for

the proposed approaches.

However the proposed method is evaluated using only one audio book and unit selection TTS, it should be tested on other books and state of the art TTS systems. The proposed embedding model uses only linguistic information and it is not designed for a specific task like TTS corpus design. It could be beneficial to use a general encoder-decoder from linguistic information to acoustic information for corpus design.

CHAPTER 5

ACOUSTIC MODEL AND CORPUS DESIGN

In the previous chapter, a phoneme embedding model based on linguistic information, that can perform well in script selection, has been proposed. The presented embedding model has not been designed or adapted for voice corpus design or speech synthesis task. There are a drawback and an advantage for this general model. The acoustic information is at least as important as linguistic information in TTS voice corpus design. The drawback of the proposed linguistic model is that it does not profit from acoustic information. So we propose to use an acoustic model which transforms linguistic information to acoustic one for TTS voice corpus design. It means the discussed linguistic embedding models in previous chapter can be replaced by an acoustic model. But on the other hand, as an asset, this general embedding model could be used for any natural language processing task such as a metric for calculating target cost in hybrid TTS systems.

This chapter is organised in two main sections. The first section is about looking for an acoustic model which can be used for TTS voice corpus design. In the second section, the relation between voice corpus design and TTS system will be studied. We will investigate how the information from the voice creation process can be useful to help a unit selection-based TTS engine.

5.1 Acoustic model for script selection

Phonemes which are linguistically similar can be uttered differently and carry different acoustic information. In this section, the linguistic embedding model will be replaced with phone level embedding models trained with acoustic information. It helps to adapt the script selection method for the TTS voice corpus design task. An acoustic model which is trained by a general speech corpus is able to predict the acoustic embedding information based on an given linguistic information. Although the acoustic

information in our script selection problem is not available before recording process, a general speech corpus can be employed to train the model in practice. In order to simulate the best case scenario, the full audio book will be used to train acoustic models.

5.1.1 Models

We propose three architectures of phone embedding model for corpus design. Figure 5.1 displays the proposed architectures for the acoustic model.

In a similar way as the methodology of sections 4.4.2 and 4.4.3, the embedded vectors resulted by acoustic models will be used by a subset selection method to design the voice corpus. Since the acoustic models provide a latent vector for a given phone, two selection methods (KMeans clustering and minimization of KLD in a greedy process) can be employed. By getting average over utterance's phones and then assigning a fixed length vector to each utterance, the KMeans clustering selects a subset of utterances. In the other approach, the embedding vectors of an utterance can be used directly for the minimization of KLD between distribution of diphones in the selected voice corpus and $\mathcal{F}$.

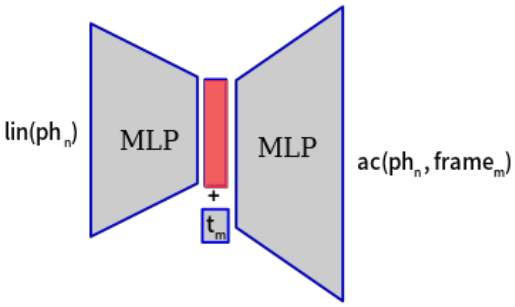
The proposed phone embedding models are described below.

MLP

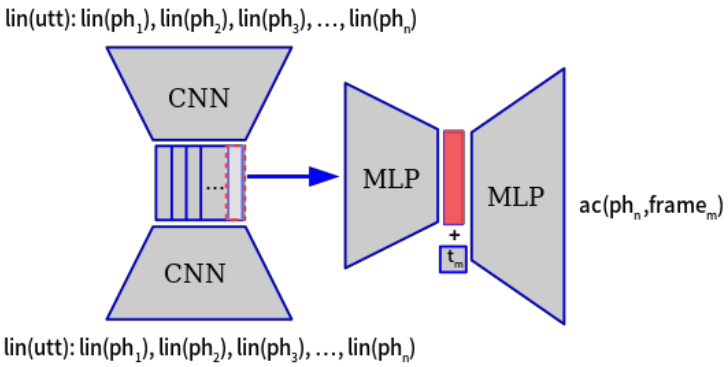
The *MLP* model works at the phone level and is displayed in figure 5.1a. It provides an embedding vector corresponding to each phone.

According to the proposition detailed in (Perquin et al., 2018), a feed-forward DNN is trained to predict the acoustic information at the frame level for each input linguistic vector. Since a given linguistic vector ($lin(ph_n)$) can correspond to several acoustic vectors (frames), the timing features of each frame are taken into account. The timing features (t_m in figure 5.1) are concatenated to embedding features (with size 30) in order to help the prediction of the corresponding acoustic features ($ac(ph_n, frame_m)$). The timing features are the phoneme duration in seconds and the relative position of the associated frame inside the phoneme.

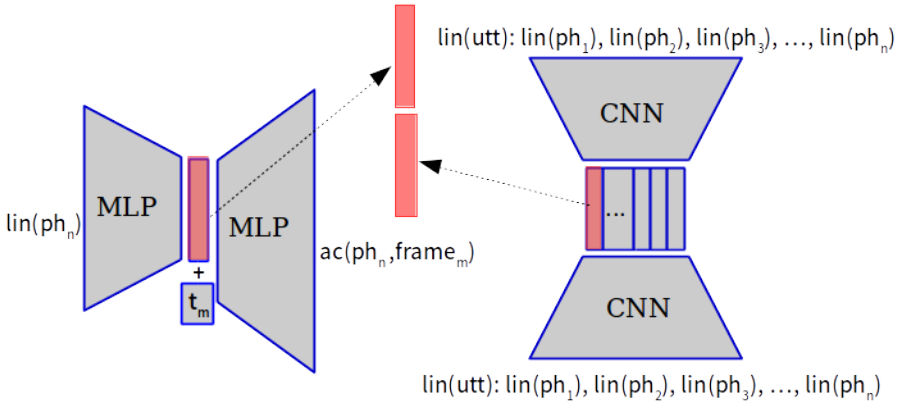
Learning data is the linguistic and acoustic information corresponding to phonemes and frames of the voice corpus $\mathcal{S}$. The acoustic features consist of a 60 dimension



(a) MLP phone embedding



(b) CNNMLP phone embedding



(c) Conc(CNN,MLP) phone embedding

Figure 5.1 – Three proposed acoustic models which provide the embedding vectors in red color.

MFCC (Mel-Frequency Cepstral Coefficients) vector, and the log of fundamental frequency F_0. The acoustic features are centered and reduced (unit variance). The frame length is 10ms.

However this model takes into account the acoustic information of utterance's phones, it does not profit from contextual information.

CNNMLP

In this proposition, the linguistic based model (*CNN*) described in section 4.3 and *MLP* model are joined sequentially (figure 5.1b).

The *CNN* auto-encoder represents the linguistic information of phonemes by a vector of latent features with a size of 30. The model is trained at the utterance level with the full corpus $\mathcal{F}$ and uses only linguistic information ($lin(utt)$). One of the assets of this model is to have contextual information of phonemes at the utterance level which could help a better representation in the embedding space. By providing utterance's phonemes information in CNN architecture, an embedding vector of linguistic features and contextual information will be produced. The linguistic embedding vector will be fed to a *MLP* model. This architecture will be called *CNNMLP* in the remainder.

Concatenation of CNN and MLP

The third proposed architecture is a concatenation of embedding vectors from *MLP* and *CNN* phoneme embedding models. In order to have maximum information, the embedding vector of each model have the same size as in previous models (30). It means the size of concatenated embedding vector is 60. This model is displayed in figure 5.1c and will be called *Conc(CNN,MLP)*.

5.1.2 Experiments and results

In the script selection problem and before the recording process, it is not possible to use the acoustic information of the considering audio book. Here we used $\mathcal{F}$ voice corpus to train all models, including the acoustic model. Consequently, this may be considered as the best possible acoustic model by using maximum information with the same context.

Acoustic model	MCD (dB)	BAP (dB)	V/UV (%)	RMSE(F_0) (Hz)
MLP	5.03	0.21	14.23	18.16
CNNMLP	5.52	0.22	35.41	0.65
The best reported in (Perquin et al., 2018)	5.06	0.35	12.6	17.9
The DNN model reported in (Wu et al., 2016)	4.54	0.36	11.38	9.57

Table 5.1 – Evaluation of acoustic models in comparison with state of the art models. The reported numbers for (Wu et al., 2016) resulted on different data. *MLP* and *CNNMLP* achieve an comparable result for predicting acoustic features.

The table 5.1 compares the proposed acoustic models with the state of the art acoustic models (Perquin et al., 2018; Wu et al., 2016). The predicted acoustic features in *Conc(CNN,MLP)* model is exactly same ac as *MLP*. We use four common measures to evaluate the acoustic models: Mel-Cepstral Distortion on MGC coefficients (MCD), distortion measure on BAPs (BAP), Voiced/unvoiced error rate (V/UV), Root mean squared error on F_0 (RMSE(F_0)). These objective measures indicate the quality of the predicted acoustic features.

The *MLP* model is same as the one in (Perquin et al., 2018) with different embedding size. Reducing the size of embedding vector from 64 to 30 reproduced almost same quality in terms of objective measures. The training data is same. Lower error in prediction of F_0 has been observed for *CNNMLP* in comparison with *MLP*. It could be explained by the benefits from contextual information which could be helpful for F_0 prediction in the case of *CNNMLP*. On the other hand, the other objective measures show slightly lower accuracy in prediction which could be explained by feature compression in the CNN auto-encoder bottleneck.

Although the main purpose of the acoustic models in this experiment is corpus design, the acoustic models in (Perquin et al., 2018; Wu et al., 2016) have been designed for synthesizing speech. It also should be taken into account that the training data in (Wu et al., 2016) was different. The voice corpus in our model and (Perquin et al., 2018) is highly expressive compared to (Wu et al., 2016). Considering this, results of objective measures can be considered as acceptable.

After the training process with the full voice corpus, the CNN model in section 4.5 is replaced with the proposed phone embedding models. The phone embedding models are followed by the KMeans clustering (see section 4.4.2) and KLD minimization (see section 4.4.3) for selecting a sub-set of $\mathcal{F}$. The sub-set voice selection is done for 5 rates out of $\mathcal{F}$ (10%, 20%, 30%, 40%, 50%). The sub-set voice corpora are used to

synthesize the test section $\mathcal{T}$ with an expert based unit selection TTS (Alain et al., 2017). Figure 5.2 compares the TTS global cost of synthetic signals. It approximate the overall quality for proposed acoustic model as the phone embedding. The *CNN-KMeans*, described in section 4.5.2, is represent our base line.

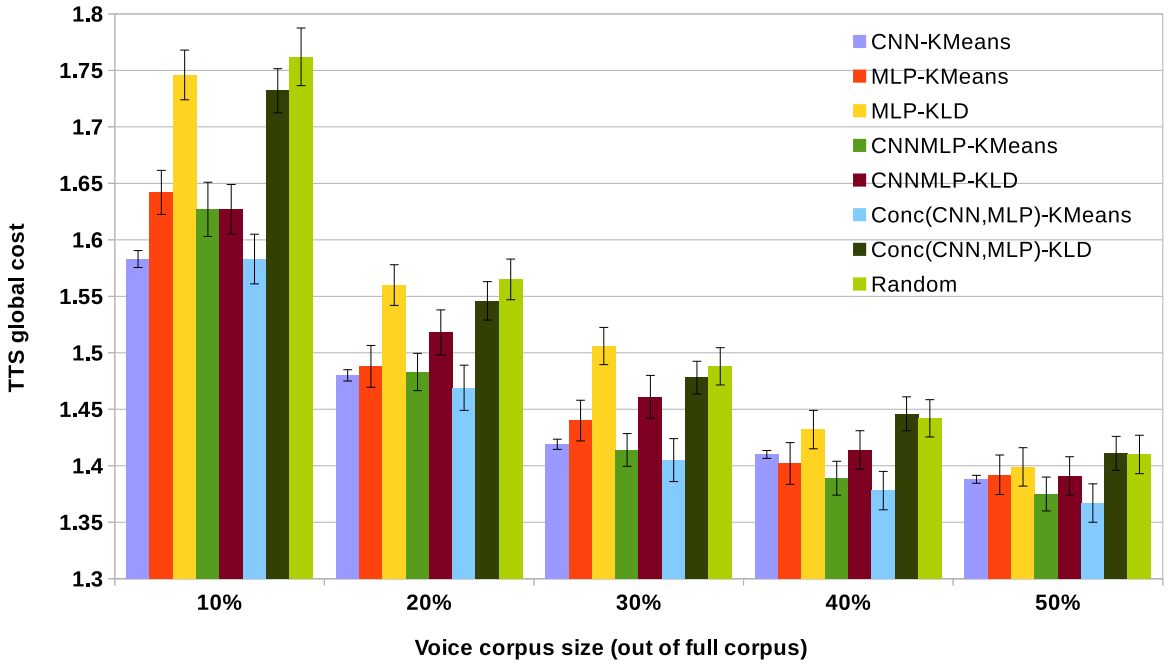


Figure 5.2 – TTS global cost resulted for synthesising test section using resulted voice subcorpus with different methods. A lower TTS cost approximates higher perceptual quality. The *CNN-KMeans* which use only linguistic information achieves lower TTS cost.

The TTS global cost shows the acoustic models, except *MLP-KLD*, perform better than the random method for corpus design. By using the *MLP* architecture followed by KMeans utterance selection, the sub-voice selected gives a lower TTS global cost than the random method for all voice corpus sizes. Despite of this, the TTS global cost does not show any increased performance of using acoustic models in comparison with linguistic phoneme embedding model (*CNN-KMeans* method). In the best case of acoustic models, *Conc(CNN,MLP)-KMeans* achieved same synthetic quality as *CNN-KMeans*. It was expected since all linguistic information in CNN auto-encoder exists in *Conc(CNN,MLP)*. It also reveals that the *MLP* embedding vector does not contain additional useful information for corpus design.

The performance of the CNN phonemes embedding for corpus design against

acoustic models evokes the idea of using this linguistic embedding in hybrid TTS. While the CNN embedding model is employed for corpus design as well, it could be expected that by using it in a hybrid TTS the synthetic signal could be improved.

5.2 Hybrid TTS using linguistic embedding model

Recent studies and the recent Blizzard challenges have revealed good achievements of hybrid systems (see for instance (Fan et al., 2014; King et al., 2018; King et al., 2017)). A unit selection based TTS is looking for most similar units to a target unit, but in voice corpus design we are looking for most different units to increase the variety. The acoustic model in hybrid TTS works as target cost and is able to find the best candidates. It means it should be able to categorize the present units in voice corpus. We are facing the same problem in the voice corpus design when the unique units are of broader interest and the similar units should be removed from voice corpus. Since the phoneme embedding model shows an impressive result in unit selection TTS voice corpus design, the performance of the proposed model in hybrid TTS system is evaluated in this section.

Beside the evaluation of the performance of the linguistic embedding model in a hybrid TTS, there is another question that we are trying to answer: *Is it helpful to use the same phoneme representation in the corpus design step and in the TTS target cost?* By comparing the proposed linguistic embedding model and an acoustic embedding model in TTS systems, the relation between TTS system and voice corpus design is investigated. It could help to improve TTS systems or technically guide the TTS voice corpus design. If the proposed linguistic embedding can be used in hybrid TTS as well as acoustic phone embedding, the importance of contextual information would be highlighted. In this case, we could provide a hybrid TTS system which does not need acoustic information for training and can be train with only script.

5.2.1 TTS systems

Three methods for calculating the TTS target cost are compared. An expert target cost function is a weighted sum of linguistic features. The two other methods are based on embedded representations in phone level. The first one uses the same embedding for the corpus design step and the target cost function while the second one uses

a specific embedding for the target cost function taking into account acoustics. The target cost is computed as the euclidean distance in the embedding space between the candidate phone and the target one. Figure 5.3 displays the three approaches compared in this study.

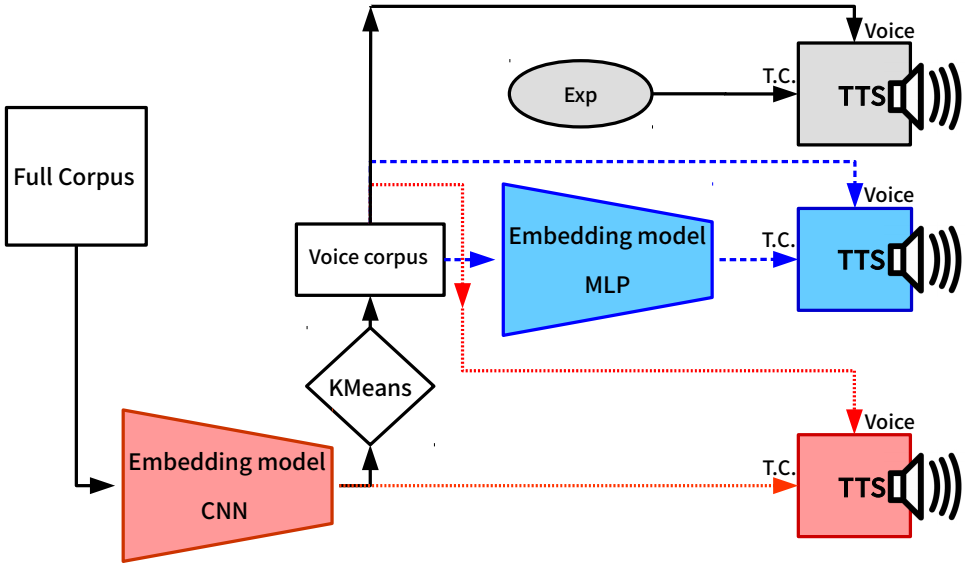


Figure 5.3 – TTS systems considered, namely Exp, MLP and CNN from top to bottom. The only difference comes from the target cost (T.C.) computation.

In the following, these three systems are described and then compared.

Expert-based target cost (*Exp*)

This expert knowledge based unit selection TTS is used as the lower band base line. The target cost is defined as a weighted sum of linguistic features and has since been improved over the years. The concatenation cost is the same as in (Alain et al., 2017), defined as a sum of euclidean distances on acoustic features between consecutive units.

Different embeddings for corpus design and TTS (*MLP*)

The second method uses an embedding model specific to the target cost function using both linguistic and acoustic information. This model has been described in section 5.1.1 and used as the state of the art unit selection TTS. After training, the encoder

part that transforms linguistic vector into embedding space is detached and used as the embedding model. The TTS target cost is the euclidean distance in the embedding space between the candidate and target units (see (Perquin et al., 2018)).

Same embedding for corpus design and TTS (*CNN*)

The third method replaces the expert target cost function by a cost function relying on the phoneme level embedding created during the corpus design step. Consequently, we propose here to use the same embedding model and phoneme representation for both corpus design and TTS target cost. As in the previous system, the target cost function corresponds to the euclidean distance in the embedding space.

Hybrid systems differences

While the *Exp* system is manually tuned by an expert knowledge, hybrid TTS systems employ an embedding model. Table 5.2 summarizes and highlights the differences of the two embedding models described above.

Method	**CNN**	**MLP**
Training data	Full corpus ($\mathcal{F}$)	Voice corpus (VC)
Input	Linguistic	Linguistic+Timing
Output	Linguistic	Acoustic
Training Level	Utterances (Sequence of phonemes)	Frames of signals

Table 5.2 – Embedding models comparison for two hybrid systems.

It is important to notice that the *MLP* model benefits from acoustics while the *CNN* model is only learnt with linguistic data. Also, both models learn, by construction, an embedding at the phoneme level, even if the *MLP* model is trained at the frame level.

The learning data of the *CNN* model are samples at the utterance level whereas the *MLP* one considers samples at the frame level. There are two advantages for the *CNN* model. First the contextual information in utterance level can be helpful for discriminating phonemes in embedding space. Second it is trained on the full corpus $\mathcal{F}$ and not only on the voice corpus VC. It provides 3339 (utterances) training samples. On the other side, the *MLP* has much more training samples (frames) with only VC corpus.

Considering all this, we want to see if the consistency of embeddings between the corpus design step and the synthesis step helps to improve synthesis.

5.2.2 Experiments and results

In order to asses the generated signals by different TTS systems, an automatic evaluation and perceptual comparisons are prepared. The TTS voice corpus is a subset of of $\mathcal{F}$ (10%) which has been extracted using *CNN-KMeans* method.

In the following subsections, we report the objective and perceptual evaluation results. In this evaluation, the *Exp* system is our lower base line and the *MLP* is the state of the art hybrid TTS.

Objective evaluation

Since for the three methods, the target cost functions measure distances in three different (embedding or not) spaces, it is not possible to compare their outputs based on TTS costs. However, the same script is used as the test set and the *Concatenation rate* is then more appropriate to compare TTS performances. For each test utterance, this is the number of concatenations in synthetic signal divided by the total number of possible concatenations. As for this measure, the lower is the better as it means more consecutive units from the same utterance. Less concatenation is assumed to result in higher quality. This measurement is computed for the test part ($\mathcal{T}$) and the rest of full corpus ($\mathcal{F} - VC$). It helps to find how methods can be generalized to other scripts than $\mathcal{F}$.

As shown in table 5.3, the *CNN* method has better statistics than *Exp* method and *MLP* beats both for test part.

Measures / Methods	*Exp*	*CNN*	*MLP*
Rest of full corpus ($\mathcal{F} - VC$)	56.63±0.16	**54.36±0.16**	**54.34±0.15**
Test part ($\mathcal{T}$)	56.64±0.52	56.24±0.51	**53.98±0.50**

Table 5.3 – Concatenation rate (%) results of synthetic signal with different TTS systems. Confidence intervals are calculated using bootstrap method with $\alpha = 0.05$.

Perceptual evaluation

In (Perquin et al., 2018), the use of an acoustic model for the derivation of target cost has proved to be superior to an expert-based model. So two AB listening tests have been prepared to compare the synthetic quality of systems. The first one is between the *Exp* method and the *CNN* method and the other one is between the *CNN* and the *MLP* method. According to the protocol proposed for perceptual evaluation in (Chevelu, Lolive, et al., 2015), each AB test is composed of the 100 samples extracted from $\mathcal{T}$ with the highest DTW on MCep features. The samples are one or several breath groups with the duration of shorter than 7 seconds.

The listeners have been asked to compare 40 pairs in terms of overall quality. The results are reported on Figure 5.4.

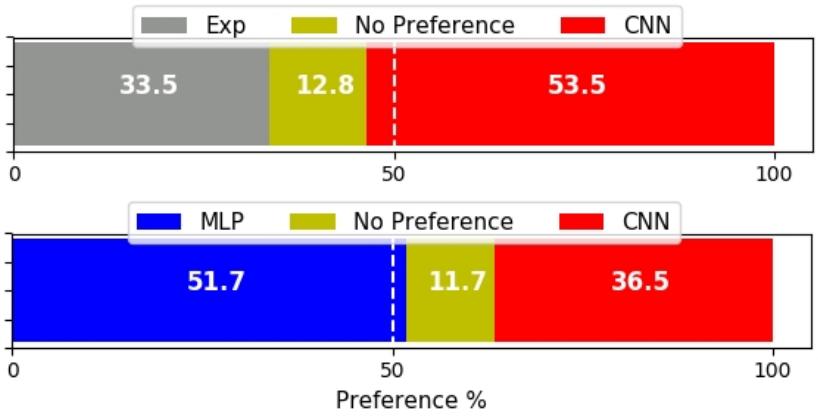


Figure 5.4 – Listening test results for comparing expert knowledg based TTS (*Exp*), hybrid TTS using linguistic model (*CNN*), and hybrid TTS using acoustic model (*MLP*).

There are 14 listeners who have participated to the first test and 10 listeners as for the second test. Each pair of samples in the first test has been compared at least 5 times and in the second test at least 4 times. The result of the first test shows that the *CNN* based embedding as input of target cost can generate synthetic signals with significantly higher quality than the expert target cost. The second test indicates the preference of listeners for *MLP* model, which takes advantage of linguistic and acoustic information, rather than *CNN* model.

This experiment has been published as a conference paper in (Shamsi et al., 2019b).

5.2.3 Conclusion

We have investigated the use of linguistic embedding model in hybrid TTS system. This embedding model which uses the phonological information, drive the TTS voice corpus as well. It can be applied instead of the expert TTS cost or an acoustic model of phones. It has then been used to build a hybrid system by computing the target cost function as the euclidean distance between units in the embedding space.

The proposed *CNN* model has been applied to provide a phoneme embedding in hybrid TTS instead of an acoustic model (*MLP*) trained on the selected voice corpus. The perceptual test has shown that the *CNN* model has better performance than expert-based target cost TTS. But the *MLP* model has been preferred to the *CNN* model which shows the importance of acoustic information.

5.3 Conclusion

However the CNN phoneme embedding model was a proper feature extractor for script selection, it was not designed for the TTS corpus design task. In the first step, we proposed three different architectures for phone embedding model which profit from acoustic information as well. These models are employed for voice corpus design. Their synthetic signals have been compared with the synthetic result of *CNN-KMeans* as the best linguistic phoneme embedding model. The TTS global cost as the objective evaluation did not show any superiority of the acoustic model in comparison with the linguistic model for corpus design. However, according to TTS global cost of synthetic signal (see figure 5.2), the acoustic models could achieve better voice corpus than random voice corpus design (except for *MLP-KLD*). Briefly, a method profits from acoustic information, which be able to improve the script selection process, has not been found.

We then investigated the relation between voice corpus design and unit selection TTS. The result showed that using an acoustic model as phone embedding model could outperform the proposed CNN phoneme embedding model as the target cost function in hybrid TTS. Although the voice corpus had been designed by CNN phoneme embedding model and the acoustic model uses less amount of data for its training. It shows that however the CNN phoneme embedding could be used for hybrid TTS system, the acoustic information are more important.

Since the acoustic information is important in TTS voice corpus design, the reasons

of a lower performance of acoustic model compared with linguistic embedding model should be investigated in future work. For instance, other acoustic models could be tested, such as using a phone duration predictor in the input of acoustic model and using state of the art acoustic models.

We also tried to implement an attention-based acoustic model (Vaswani et al., 2017) to predict a sequence of frames based on the sequence of phones information. With this model, we observed that the prediction of frames at utterance length or even breath group length is a difficult task for this model. The input of this model is a given phone information and output is the sequence of corresponding frames. It could be the reason of high error rate in the reconstruction of acoustic features. Using a phone duration predictor of an input can be tested in future works.

While the result emphasise on the importance of acoustic information, we will analyse the subcorpus achieved by the best script selection method. It helps to find out if there is any features that are of interest for TTS voice corpus design.

CHAPTER 6

SHORTEST UTTERANCES

Some studies point out that the reduction algorithms have a tendency to select shorter utterances, in a less or more important manner according to the initial corpora, reduction criteria and algorithms, as it can be observed in (Barbot et al., 2015) for instance. In (François et al., 2001), the set covering method selects a sub-corpus with an average length of 20 phonemes per sentence out of an initial corpus with an average length of 74. It is also noticed in (Cadic et al., 2010) where authors proposed to correct the algorithm to force longer sentences. A same trend has been observed in (Van Niekerk et al., 2017) in which KLD selection algorithm based on diphone distribution has preferred shorter sentences. In these cases, it may be explained by the expert function that is optimized locally by the greedy algorithms. It would then be a bias of the algorithms and not a trend from the data to achieve better quality.

On the other hand, the *CNN_Utt_KMeans* (will be called CNN-KMeans) method in the previous experiment is completely unsupervised but it also selects shorter utterances nonetheless. We can illustrate this with the measures reported on figure 6.1. It displays the average length – in number of phonemes per utterance – of the sub-corpus built at various reduction rates and from two different algorithms. We can observe that the unsupervised CNN-KMeans system selects significantly shorter utterances. Moreover, the higher the reduction rate is, the more constrained the optimisation problem is and the stronger the trend to select shorter utterances is. This trend is observed for both algorithms.

These observations lead us to question ourselves about the impact of the length of the selected sentences on the final TTS quality: is it a consequence of the optimization or a cause of the good results? Let us assume that it is a cause and name it the "shortest" hypothesis. If a voice created by selecting only the shortest sentences is less good than a voice from another strategy, it will allow us to discard this hypothesis. On the contrary, if all attempts show that the "shortest" strategy is better, it will give us clues that this hypothesis may be true and encourage us to investigate further.

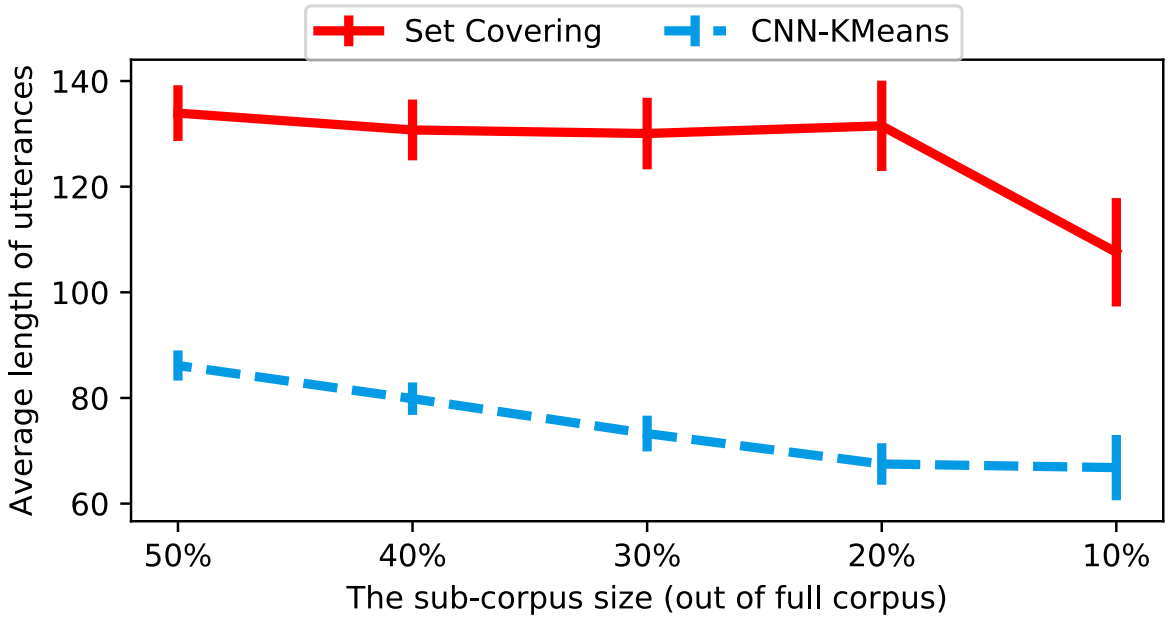


Figure 6.1 – Average sentence length of sub-corpora provided by two reduction algorithms and at various reduction rates of the *Pod* corpus (see section 6.1 for details on algorithms and corpus). The best system selects shorter utterances. Besides, increasing the parsimony size constraint involves a decreasing of the length of the selected sentences.

In this chapter, we will test this "shortest" hypothesis by simulating a voice creation process on different kinds of books (one with long formal sentences, recorded by a male speaker and one with shorter and less formal sentences, recorded by a female speaker) with different kind of TTS engines (one expert unit selection TTS and one hybrid TTS). We will compare one of the best reduction strategy proposed so far, i.e. the CNN-KMeans from previous experiment, with a simple "shortest first" algorithm using automatic measures and perceptual evaluations. Then we will investigate if the classical optimisation criteria – linguistic unit coverage or distribution – can predict or explain the observed results.

This work has been published as a conference paper in (Shamsi, Chevelu, et al., 2020).

6.1 Data and systems

The methodology and the materials of the experiment will be explained in this section. The voice corpus design algorithms, two different audio books as corpora, and two unit selection-based TTS systems will be detailed in the following.

6.1.1 Script selection algorithms

To simulate the corpus design process by reducing a full corpus, four approaches are considered below.

Random

This simple baseline consists in selecting a sub-set randomly until the requested length is reached. Since this approach is less stable by design, statistics resulting from this method and detailed further are consolidated by repeating this selection process six times. Each utterance in the *test section* will also be synthesized six times and the associated average score will be taken into account for the objective evaluations.

Set covering

A greedy based approach is used here, as presented in 4.5.2. The attributes considered for the coverage are the diphone labels enhanced with 20 linguistic features. Those linguistic features are Boolean variables answering questions like *"it is or not the first/second phone, in the first/last syllable?"*.

CNN-KMeans

This approach employs an embedding representation of several linguistic features to characterize utterances. The embedding space is produced by a multi-layer CNN auto-encoder implemented to project the discrete features into a continuous space. Then, for each utterance, the average vector of its embedded unit sequence is computed and used as its representation. A KMeans algorithm clusters utterances and for each cluster, the closest utterance to its center is selected.

Shortest

As presented in the introduction, to synthesize an expressive text like a book, our assumption is to use the shortest utterances first. To assert it, we propose a system named *Shortest*. Its algorithm is basically a simple loop that selects the shortest utterance until the desired length of the selected sub-corpus is reached.

6.1.2 Corpora

Two French audio books are used as initial corpora for experiments. The first one is *Pod* corpus which is introduced in 3.2.1. The second one is *Nad* corpus which is *La Vampire* by Paul Féval (Sini et al., 2018). While *Pod* contains long formal utterances, *Nad* contains more contemporary content with simpler utterances. The average length of utterances in *Nad* is less than half the one in *Pod*. Their main properties are summarized in table 6.1.

Corpus	**Pod**	**Nad**
Speaker gender	Male	Female
Number of utterances	3339	6032
Average length of utterances	120.1$\pm$3.2	54.4$\pm$1.2
Duration	10h 44min	10h 02min
Number of distinct diph.	1005	1000
Number of distinct triph.	12655	4693

Table 6.1 – The initial voice corpora details

6.1.3 TTS engines

Two types of TTS systems are used for synthesis in our experiments.

The first one, is a standard unit selection engine (Alain et al., 2017) with a beam search algorithm. The global cost function optimised by the TTS is a weighted sum between a concatenation and a target cost. The concatenation cost is a weighted distance between some acoustic features (MFCC, F0, amplitude, etc.). The target cost is a weighted distance between linguistic features (phoneme, syllable, positioning information, etc.). In this system, all weights were manually tuned over time. It then will be called *expert* TTS in the remainder.

More recently, most unit selection systems shifted to an hybrid architecture that includes DNN to learn the cost functions (King et al., 2018). Following this trend, the second system for the experiments, called *hybrid* TTS, is inspired by (Perquin et al., 2018). Its target cost is computed based on an euclidean distance in an embedding space. This embedding is learned from an encoder-decoder trained on the voice.

In the following experiments, the *hybrid* TTS uses only one DNN per speaker to compute the target cost. From our experience, the bias it may introduce is not significant and it allows to directly compare all costs between sub-voices from the same corpus. It also helps to discard noise from the DNN training initialisation.

6.2 Experimental setup

Two audio books with almost same length (around 10 hours) are provided as the initial corpora. A 10-fold cross validation without shuffling is used for separating the full corpus (90%) and test section (10%). Each fold is continuous, like a chapter, and the first fold starts with the first utterance in the book. Finally, the initial corpora will be synthesized by different full corpora and sub-corpora. The length of the selected sub-corpus is fixed to 10% of full corpus (about 1 hour).

The remainder of this section will describe the objective measures which are used to approximate the quality of sub-corpora and the synthetic quality.

6.2.1 Objective measures

It is inevitable to ask listeners for comparing the quality of the synthetic signals but listening tests are costly and need enough listeners. Based on the result in section 3.2.2, we propose to use TTS costs as the objective measures to approximate the quality of synthetic signals.

The global cost and concatenation cost of the synthetic signal of test section utterances are normalized by their length. These normalized costs average over utterances are used to compare different corpus design methods for each TTS/corpus.

6.2.2 Perceptual evaluation

The synthetic signals resulting from *CNN-KMeans* as the state of the art method are compared with the *Shortest* method ones.

By running 10-fold cross validation a sub-corpus is extracted from each full corpus to synthesize the corresponding test section for each fold. It will provide a synthetic signal of the whole book. As mentioned in table 6.1, there are 9371 utterances in the two corpora. However, the excessive length of some utterances may be problematic for listeners to compare the signals. For instance, the longest sentence in *Pod* is 238 words long (82 seconds). Consequently, each utterance is split into breath groups. Breath groups shorter than a minimum length (20 phonemes) are merged with the following breath group. It provides 37711 breath groups that have been synthesized according to the corresponding selected voice using *hybrid* and *expert* TTS for each fold of the cross validation. Based on the idea of (Chevelu, Lolive, et al., 2015), to avoid smoothing the results, pairs of signals that are too similar (DTW < 1.0) have been removed. Then, 100 sample pairs have been selected randomly from remaining candidates as the listening test samples. Half of these samples has been selected from *Pod* corpus and half from *Nad* corpus. Listeners evaluate 40 pairs of synthetic signals on a 5 points scale. At each step of the test, the script of the full utterance corresponding to the signal is displayed, even if the signal is only a part of the utterance. The pronounced part is highlighted to help listeners evaluate the overall quality of samples by considering the context.

6.3 Results

Methods mentioned in section 6.1 have been run using 10-fold cross validation to select 10% of the full corpus. The average length of selected utterances by the different selection methods are compared in table 6.2. In French, the average length of sentences depends on the context. For instance, the average length of sentences in *Le Monde*, whose context is French newspaper, is around 98 phones (Larnel et al., 1991). This length for the *SynPaFlex* corpus, which contains novel books and poems, is 48 phones (Sini et al., 2018).

Corpus	Pod	Nad
Full corpus	120.1±3.2	54.4±1.2
Random	121.1±5.0	54.9±1.8
Set covering	163.2±10.3	95.0±4.7
CNN_KMeans	86.7±4.5	38.6±1.5
Shortest	44.5±1.2	22.0±0.5

Table 6.2 – The average length (number of phones) of selected utterances for 10% of full corpus

6.3.1 Objective measures

The selected sub-corpus voices have been used to synthesize the test section of the 10 folds. The average global cost normalized by length (number of phones) of synthetic signals is shown in figure 6.2. Given that the same behavior is observed with the concatenation cost, it is not shown here.

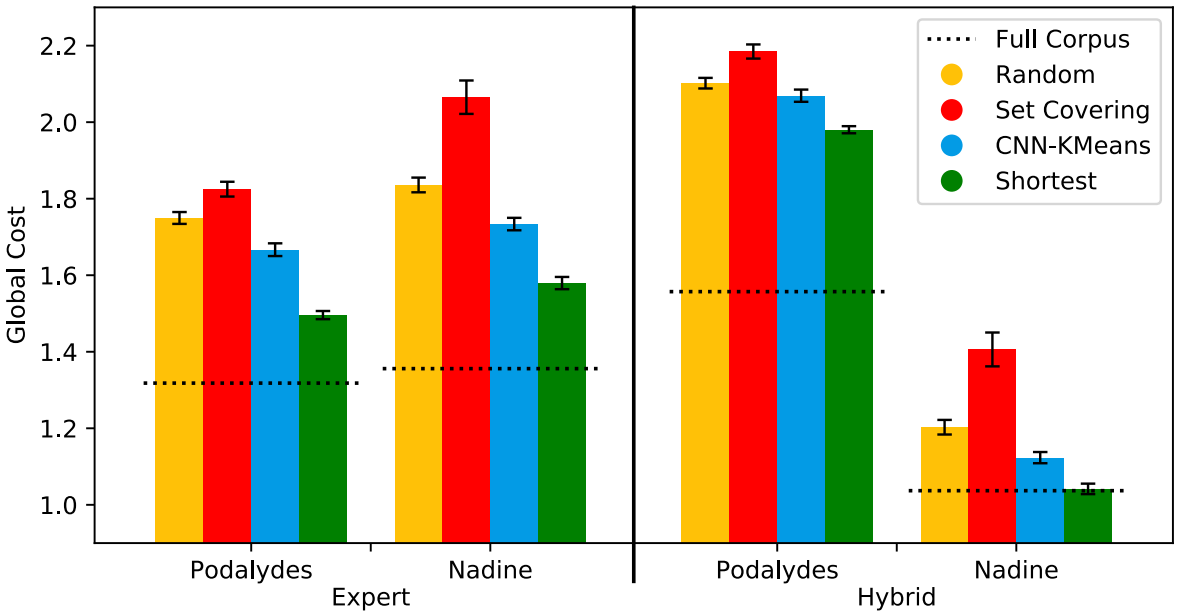


Figure 6.2 – Average TTS global cost per phone after a 10-fold cross validation. *Shortest* gives the best results in all cases.

The resulting voices from the *Shortest* method succeeds to synthesize signals with lowest global costs. The resulting signals from the *SC* method have higher global costs

than even *random* method. It shows that however following a set covering strategy will guarantee all units to be covered, the resulting TTS costs would be worst than random method for big enough voice corpora.

The voice corpus built with short utterances are expected to be less efficient for synthesizing long utterances (Kominek et al., 2003). To investigate this assumption, the correlation coefficients between the length of utterances and the TTS costs of the corresponding synthetic signals have been calculated. The Pearson correlation coefficients for the global and concatenation cost of both TTS are less than 0.12. This means even by selecting short utterances for voice corpus, TTS systems are able to synthesise long utterances almost with same cost.

6.3.2 Perceptual evaluation

Based on the TTS cost results, an AB preference test has been conducted to compare two best corpus design methods. 200 synthetic signals have been selected from the *Shortest* and the *CNN-KMeans* methods. For each combination of TTS and book, 50 signals have been chosen as the perceptual test samples.

In total, 12 listeners have compared pairs of synthetic signals. Each pair has been evaluated at least 2 times. Results are shown in figure 6.3. The perceptual results confirm the results obtained with the TTS costs and the superiority of the *Shortest* method for both corpora and TTS systems.

6.4 Analysis

In this section we look into the linguistic characteristics of result sub-corpus. First the coverage rate and distribution similarity of resulted sub-corpus by different methods will be compared. Next, the properties of short utterances will be discussed.

6.4.1 Coverage rate and distribution similarity

Other measures need to be considered to evaluate the selection method, such as coverage rate of units, or distribution similarity of units with a target distribution. The first one, the coverage rate, is defined as the number of distinct diphones/triphones which exist in the selected sub-corpus per total number of distinct diphones/triphones

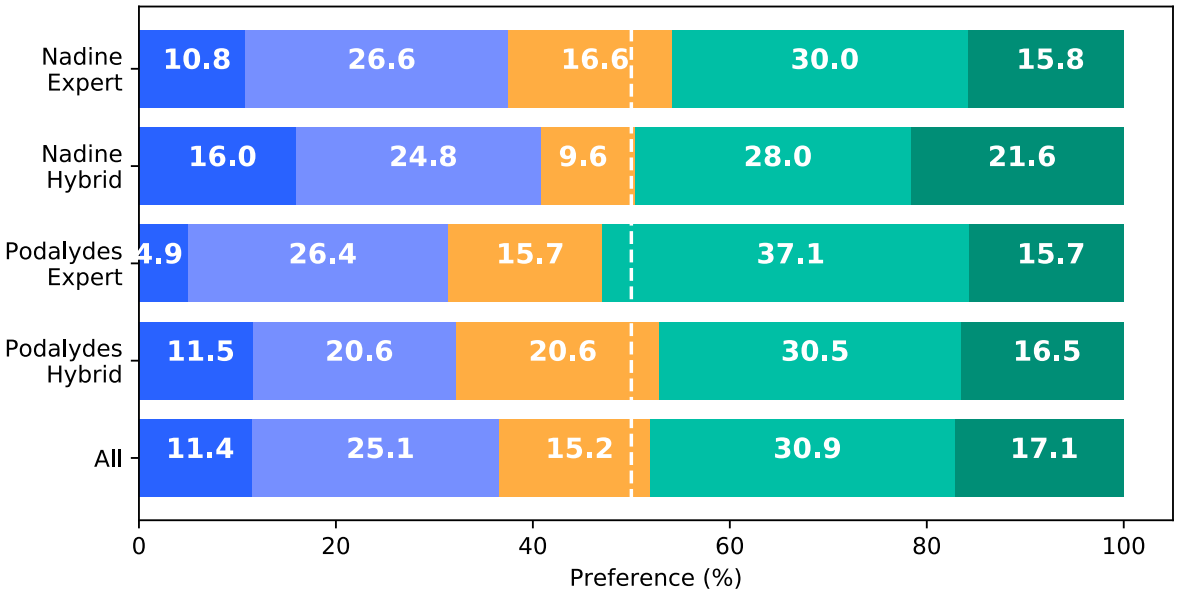


Figure 6.3 – The Perceptual test results. Right to left: strongly *CNN-KMeans* (dark blue), slightly *CNN-KMeans* (light blue), no preference (yellow), strongly *Shortest* (light green), strongly *Shortest* (dark green).

in the full corpus. The second one, the distribution similarity of diphones/triphones in the sub-corpus with the full corpus, is evaluated using KLD. The KLD indicates the dissimilarity between two distributions. Some studies claim that a lower KLD with a target distribution will result in better sub-corpora (Krul et al., 2006; Shinohara, 2014).

Figure 6.4 compares the coverage rate and distribution similarity of four methods. The top figure is the diphones and triphones coverage rate in selected sub-corpus by the different methods for the two corpora. The bottom figure is the KLD between diphone/triphone distribution of sub-corpus and the full corpus. As the KLD value decreases, the selected sub-corpus distribution is increasingly similar to the one of the full corpus. Each color circle indicates a selected part from one fold of the full corpus.

Based on table 6.1, however the number of distinct diphones in two corpora are similar, the number of the distinct triphones in *Pod* corpus is almost three times higher than in *Nad* corpus. The coverage rate of the *Shortest* method is almost same as *CNN-KMeans* and *Random* methods. It means the short utterances does not contain a set of specific units and they are as good as random in terms of unit coverage for 1 hour of sub-corpus. However the diphone coverage of *Nad* corpus with the *Shortest*

method is slightly lower than others.

Based on Figure 6.4b, it could be observed that the *Shortest* method does not respect the general distribution of corpora. While the random selection method achieves the lowest KLD, the *Shortest* method results in the highest KLD in both corpora. It is not surprising to have the same distribution as full corpus by the random selection.

6.4.2 Properties of short utterances

As it is mentioned in (Charfuelan et al., 2012; Kominek et al., 2003), short utterances often are more expressive and have a different prosodic delivery. In contrary to (Braunschweiler et al., 2011), the main idea in *Shortest* method is to have more possible prosodic variation in the voice corpus.

However the *Shortest* method can not guarantee the coverage of all diphones or phones, we hope the sub-corpus length is long enough to cover all needed phones. The alternative solution would be replacing the not selected shortest utterance which contain the missed phones with the longest utterances in the selected sub-corpus.

Needless to mention that the short utterances are easy to read in the recording process. A drawback is that the *Shortest* method will select repetitive sentences. However, in term of linguistic information, same utterances do not add new units to corpus, they can contain different acoustic information. For example, there are 5 utterances with same script (*"Ah"*) but they are completely different in terms of intonation.

As a first investigation, we find more variation of F0 in the voice corpus obtained with the *Shortest* method than others. It emphasizes the importance of acoustic and prosodic variation of the sub-corpora containing short utterances.

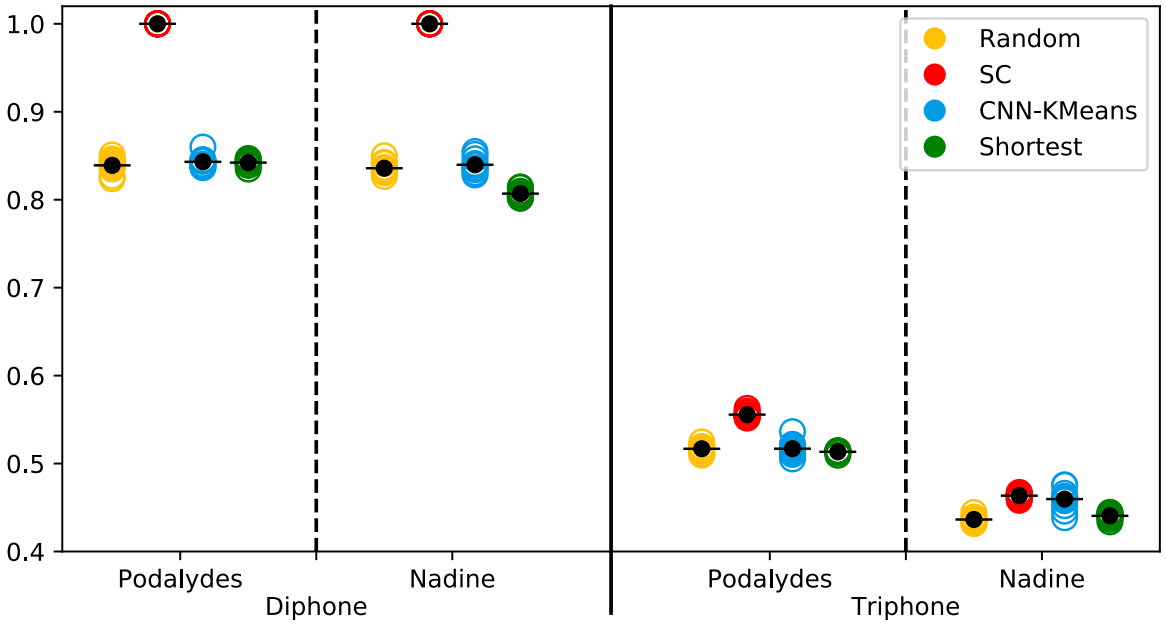
6.5 Conclusion

In this chapter four methods for TTS voice corpus design have been compared. These methods are evaluated with two kinds of TTS and for synthesizing two french audio books. The synthetic signals obtained these methods have been compared objectively using TTS costs and perceptually by listeners.

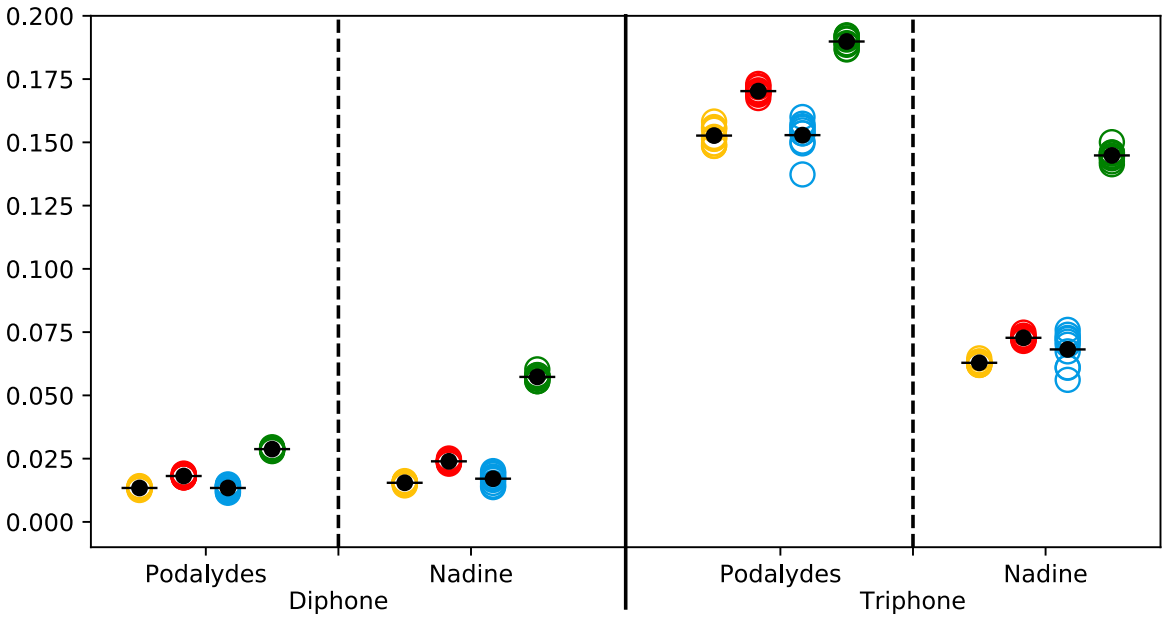
The experimental results showed a simple method like selecting short utterances could work well for TTS corpus design in audio-book generation when the voice corpus is a portion of the book. This method worked better than *CNN-KMeans* method in

hybrid and *expert* TTS for audio book with long and short utterances. The results showed that the coverage of units as the classical method does not work even as good as random selection in a large enough voice corpus. By comparing the TTS cost and the coverage rate and KLD as unit distribution similarity, it revealed that the previous strategies of corpus design (Barbot et al., 2015; Krul et al., 2006) does not lead to the best voice corpus. They are not necessarily a good metric of corpus design for big enough voice corpora in TTS.

The results and the performance of the *Shortest* method should be tested with more corpora with different average utterance length. As future work, a combined method can be proposed which takes into account the average length of utterances in book. In other words, it could be more efficient to adapt the selection process to the context and the characteristics of book.



(a) Coverage rate (higher is better).



(b) KLD (lower is more similar to target distribution).

Figure 6.4 – Coverage rate and KLD of diphone/triphone for 10-fold cross validation. Average of each column is in black.

CHAPTER 7

EVALUATION OF MIXED SYNTHETIC AND RECORDED SIGNALS

The problem of audio book generation has been described in section 2.1 and approached from the perspective of vocalization combining natural human voice and synthetic one. Selecting a recorded portion which will be used as the TTS voice corpus for synthesizing the rest of the book has been addressed in previous chapters.

The idea of mixing synthetic and human voice signals is not new in the literature. Previous studies (Clark et al., 2019; Gong et al., 2003; Lewis et al., 2006) have agreed that listeners prefer fully synthetic signals rather than signals combining TTS and natural voice. As for (Gong et al., 2003), it has been observed that users' liking and clarity of fully synthetic signals are higher than mixed TTS-human signals. In (Lewis et al., 2006), the preference of listeners is asked about mixed synthetic and recorded voice when dynamic part of a message is synthesized by TTS. Participants indicated that they preferred the fully synthetic signals. Recently authors of (Clark et al., 2019) have investigated the naturalness of synthetic sentences in three different ways: isolated sentence, full paragraph, context-stimulus pairs. Their experimental results have showed that two successive synthetic signals get higher score than a sequence composed of a natural signal and a synthetic one. Furthermore, considering some aspects of synthetic speech quality such as intelligibility, (Wester et al., 2016) has found out synthetic speech by produced recent TTS systems could be as good as human voice.

Audio-book generation is different and more challenging than tasks done in (Clark et al., 2019; Gong et al., 2003; Lewis et al., 2006) that do not require expressiveness (news or message reading). Even advanced TTS systems are not as good as professional speakers for generating expressive books in terms of overall quality.

The sub-set selection problem has been investigated by taking into account the richness of voice corpus and synthetic quality. It could be expected that synthetic signals would have less overall quality than recorded signals of professional speakers.

Regardless of the signal quality achieved by a TTS and its voice corpus, the length of synthetic portion in final audio book has been the only constraint in the audio-book generation problem considered in the previous chapters. Although the order or configuration of synthetic and recorded speech signals could be important too.

In this chapter, we investigate the configuration of mixed signals in expressive audio-book generation using a hybrid TTS system. There are two main motivations for the experiments conducted in this chapter. First, the initial idea of audio-book generation as mixed signals will be examined by comparing fully synthetic signals to mixed synthetic and natural signals based on their perceptual quality. In other words, the main aim of this study is to answer these questions: *In terms of overall quality, is it helpful to generate an audio book with mixed synthetic and natural signals? Or do listeners prefer a fully synthetic audio book?* Second, in the case of preference for mixed signals, the impact of the order of synthetic and natural signals on the perceived quality will be investigated. These experiments will be done considering different levels of synthetic quality.

This experiment has been published as conference papers in (Shamsi, Barbot, et al., 2020).

7.1 Perceptual evaluation

In (Hinterleitner et al., 2011), a protocol for subjective evaluation of TTS in audio book reading tasks has been presented. The authors have suggested asking listeners to assess the quality of an audio book using 11 criteria such as listening pleasure, listening effort, intonation, emotion, etc. We believe that these terms are not always clear and do not have common definitions among listeners. Since the target of audio-book generation task is ordinary people, not expert voice quality annotators, we suggest asking for overall quality or overall preference of listeners.

The goal of the first experiment is to compare the overall quality of fully natural speech, fully synthetic speech, and a mix of natural and synthetic signals for expressive audio-book generation using a hybrid TTS system. The experimental framework of studies previously cited was completely different from the one considered here. For example, the naturalness of synthetic signals generated by a vocoder-based TTS in (Clark et al., 2019) or synthetic quality of an HMM-based TTS in a non-expressive context in (Gong et al., 2003) have been evaluated. Considering our objective, it seems

without transition	1 transition	2 transitions	3 transitions
NNNN	NNSS	NSSN	NSNS
SSSS	SSNN	SNNS	SNSN

Table 7.1 – Different transition configurations for 4 parts (breath groups). Natural parts are indicated by N and synthetic parts by S.

to be necessary to evaluate again the hypothesis of listeners' preference for fully synthetic signals and mixed signals. Moreover, mixed audio-book generation, which has been described in section 2.1, needs to take into account the impact of signal type order (synthetic or recorded first) in overall quality of the final audio book.

The quality of synthetic signals using different TTS settings may vary, especially in expressive tasks. In order to take into account this variability, different quality levels of synthetic signals have been considered in this study.We have observed that TTS voice corpus size has a direct impact on synthetic quality (see 3.2.2). Consequently, we propose to conduct all perceptual tests by using synthetic speech built from 3 voice corpus sizes (30 minutes, 1 hour, 5 hours).

Regardless of the quality degradation of synthetic signals in comparison with recorded speech, the change of signal type in a mixed signal sequence could be disturbing for audio book listeners. We call this change a transition. Transitions can happen from synthetic speech to recorded natural speech or the contrary. In order to examine their impact on overall perceptual quality, 8 different configurations are evaluated (see table 7.1). In total, each sample is prepared with these 3 voice corpus sizes and these 8 transition configurations.

Some studies such as (Chiaráin et al., 2017; Latorre et al., 2014) have emphasized on the importance of context in voice perception. For instance, in (Latorre et al., 2014), it has been found that, without context, listeners do not always prefer the signals produced by humans. It leads to evaluate the speech perception using long-form speech. A possible drawback of this approach is the exhausting nature for listeners to assess an entire chapter or even several paragraphs of an audio book and thus a reduction of the evaluation reliability. We propose then four consecutive breath groups to construct perceptual test samples (a breath group instead of synthetic/recorded part in table 7.1). The long (more than 70 phones) and short (less than 45 phones) breath groups are filtered out from the candidate list for listening test samples. The average duration of breath groups are 3.49 ± 0.40 seconds. It helps listening test samples to have reason-

able duration (around 14 seconds) and containing almost same synthetic and recorded lengths. In order to provide some context, the transcription of the signal plus the script of the utterance just before and after the test sample are provided to testers.

7.2 Experiments and results

From the *Pod* corpus (see section 3.2.1), three voice corpora (30 minutes, 1 hour and 5 hours) are randomly extracted.The smaller corpora are included in the larger ones. As for speech synthesis, the hybrid TTS described in section 6.1.3 is used.

The listening test transcriptions are extracted from the rest of the book which is not selected for voice corpora according the following methodology. The sequences of four consecutive breath groups with a duration between 3 and 6 seconds are listed. They are not limited to only one utterance and can belong to several consecutive utterances. Out of this list, 20 transcriptions (80 breath groups) have been selected randomly for the listening test samples. These breath groups are synthesized using the three voice corpora. Each configuration of mixed signals in table 7.1 is prepared by using synthetic and natural signals.

7.2.1 MOS test

A MOS test is designed for evaluation. Listeners are asked to rate the overall quality of each sample on a scale from 1 to 5 with a step of 0.5. The cognitive load of a long perceptual test causes unreliability of evaluation. Consequently, in order to keep the quality of evaluation, only 25 samples are provided to each listener which takes around 12 minutes to be evaluated. In the following section, the result of this perceptual evaluation is presented.

7.2.2 Result

In total, 29 non-expert listeners participated to the evaluation which gives 725 scores. Table 7.2 details the main results of this perceptual test and the confidence intervals of score average calculated using the bootstrap method with $\alpha = 0.05$.

These MOS scores are not comparable between different languages and test settings. The average score for human voice (4.32) is lower than in previous studies (Clark

Num. of transitions	Config.	5 hours	1 hour	30 minutes
Fully synthetic	SSSS	2.70±0.41	1.64±0.36	1.31±0.39
1 transition	NNSS	3.67±0.28	3.08±0.38	2.86±0.30
	SSNN	2.90±0.39	2.30±0.38	1.81±0.41
2 transitions	NSSN	3.30±0.33	3.03±0.26	2.20±0.38
	SNNS	3.30±0.34	2.87±0.35	2.58±0.41
3 transitions	NSNS	3.48±0.31	2.74±0.29	2.67±0.42
	SNSN	3.68±0.31	2.62±0.34	2.42±0.41
Human voice	NNNN	4.32±0.17		

Table 7.2 – MOS test results for evaluating mixed synthetic/natural signals. Mixed signals are evaluated with higher quality than fully synthetic signals.

et al., 2019) (around 4.6). Moreover, the question evaluated in our experiment is the overall quality while the MOS score in (Clark et al., 2019) corresponds to naturalness.

Based on these results, the mixed signals have significantly higher scores than fully synthetic signals in all voice corpus sizes. This observation is contrary to the previous studies (Clark et al., 2019; Gong et al., 2003; Lewis et al., 2006) which showed superiority of fully synthetic signals. On the other hand, figure 7.1 does not show any significant difference when the number of transitions changes. However, mixed signals with 3 transitions have slightly higher score in comparison with others in 30 minutes and 5 hours voice corpora. In case of long synthetic part (*SSNN*, *NNSS*, *SSNN*), the following ranking between MOS scores can also be observed: *NNSS* $\geq$ *NSSN* $\geq$ *SSNN*.

7.2.3 Preference test

In order to investigate more the impact of transitions, another perceptual test is proposed. The configurations given in table 7.1 are categorized into two groups: the first one corresponds to configurations with a *long* synthetic part and the second one to the *short* synthetic part configurations (*SNNS*, *SNSN*, *NSNS*).

We propose to use a simple protocol, an AB test, to directly compare the *long* and *short* categories. This test is designed with 3 levels of preference (no preference, slightly, strongly). For a same transcription and a same TTS voice corpus size, a sample with *long* synthetic part is compared with a sample stemming from the *short* category. The 3 voice corpus sizes are considered. Listeners are asked to compare

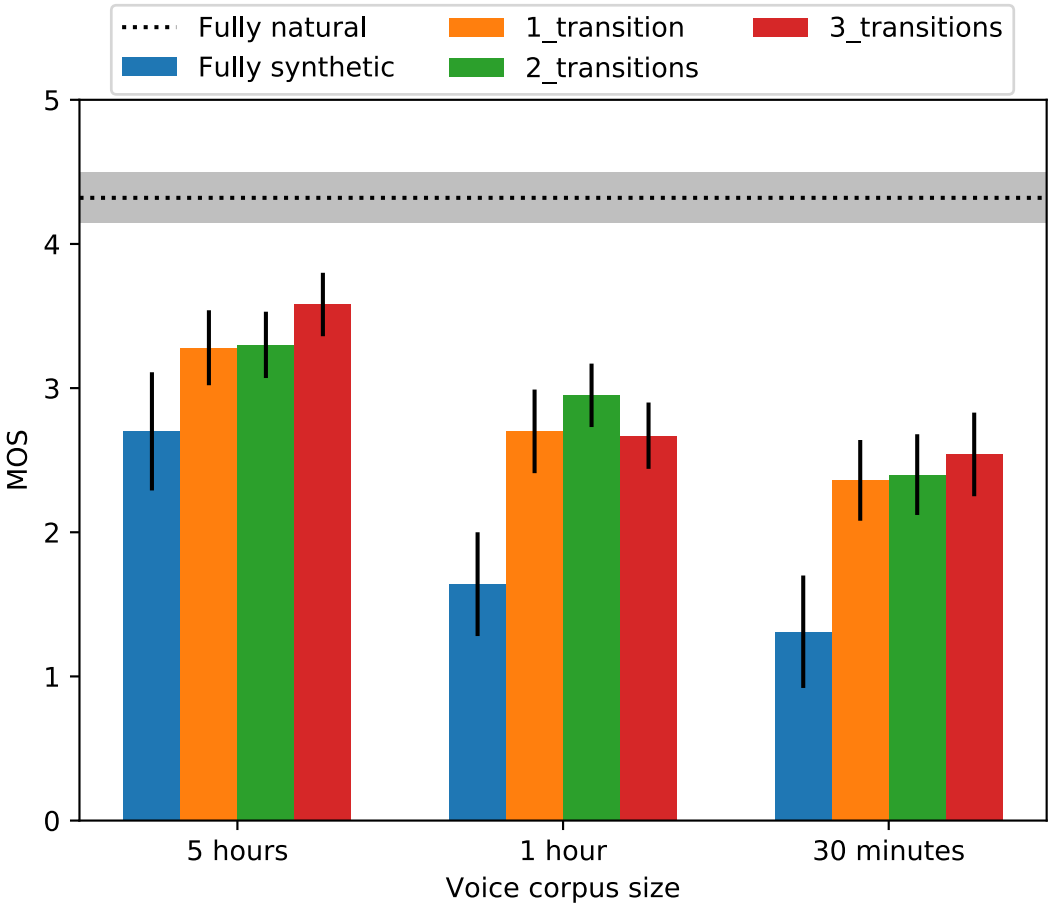


Figure 7.1 – MOS test results for evaluating mixed synthetic/natural signals (aggregated based on voice corpus size)

25 pairs of samples which takes around 25 minutes time.

26 listeners have done the test, which resulted in 595 comparisons. Results are shown in figure 7.2.

The result does not reveal any significant difference between *long* and *short* synthetic part in mixed signals. According to listeners' feedback, sometimes the comparison is very difficult. Despite of this, listeners had no preference between samples only 20.1% of times.

If we remove comparisons of pairs with 2 transitions (*NSSN* and *SNNS*), the AB test changes to a direct comparison between one transition and three transitions. In this case, the preference results do not show any difference between those too configurations.

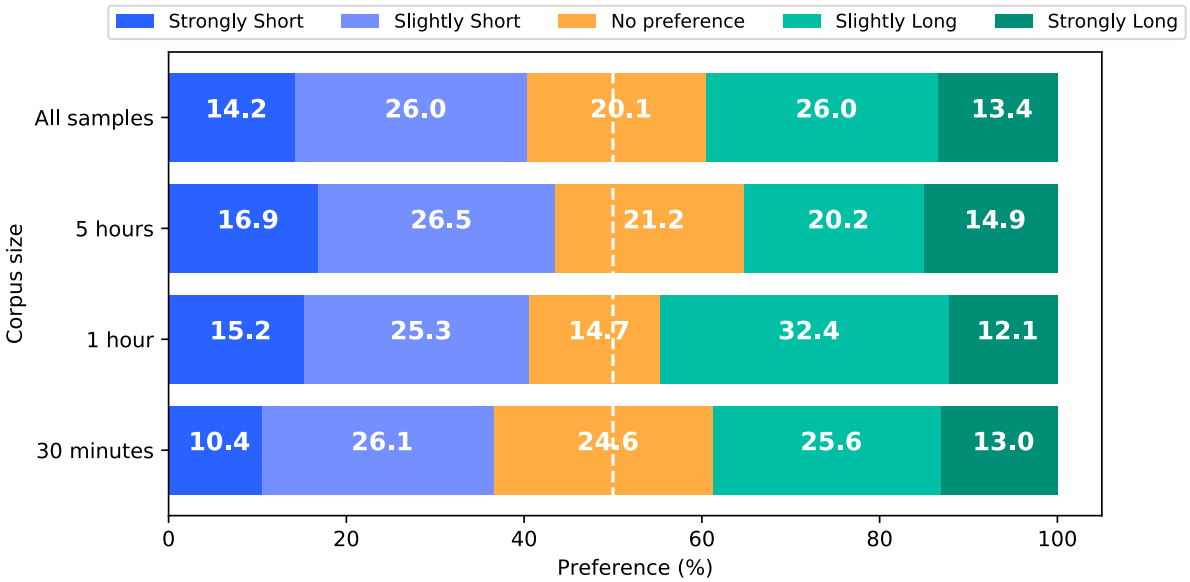


Figure 7.2 – AB test result for evaluation the impact of the length of continuous synthetic part in mixed signals.

7.3 Results analysis

Two considerations about perceptual test results will be followed. First the result of the perceptual test will be considered based on signal quality instead of the size of voice corpus. The initial idea of using different sizes of voice corpus was to simulate different synthetic quality levels. After assessing perceptual quality of fully synthetic signals, following this simulation is not necessary. Consequently, the resulting perceptual score will be categorized based on the resulting perceptual quality of fully synthetic signals instead of voice corpus size. Afterward the impact of starting and ending parts will be investigated.

7.3.1 Investigate of synthetic quality

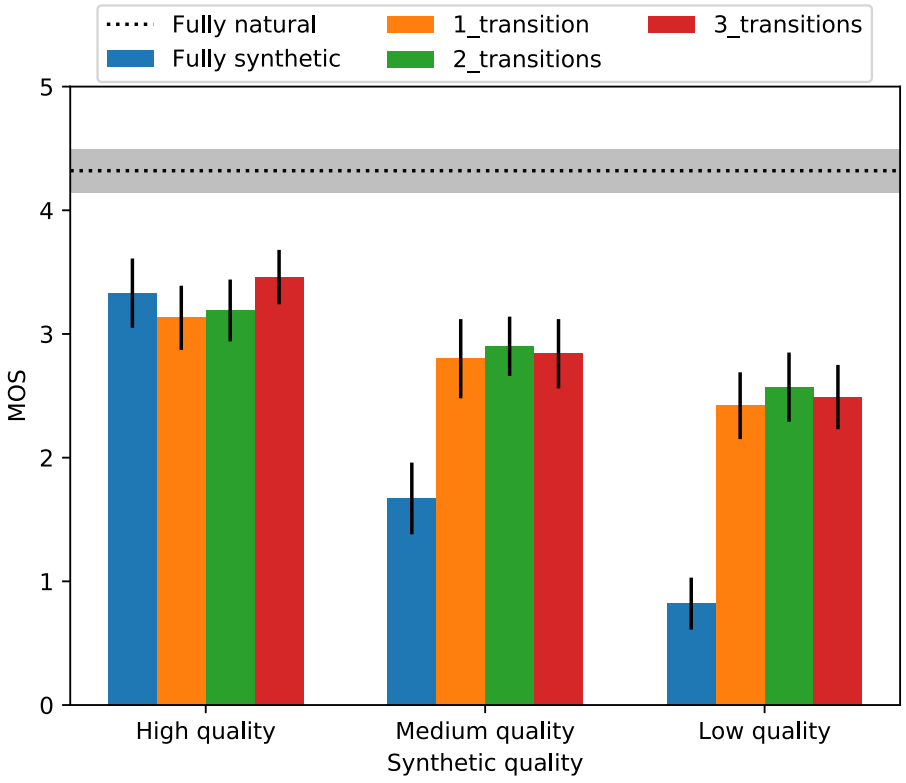
The inverse configuration of mixed sample, e.g. SNSN and NSNS, in listening test are existed in samples set. So the average length of synthetic part in mixed signal is same as the average length of natural part. But the length of synthetic parts has variations among listening test samples. It could be claimed that samples with longer synthetic parts would be evaluated with small MOS scores. To examine this hypothesis

we calculate the correlation coefficient of MOS scores and synthetic part lengths after removing *NNNN* and *SSSS*. A low correlation (Pearson: -0.20, Spearman ranking correlation: -0.21) rejects the relation between MOS score and synthetic part length. On the other hand the Pearson correlation coefficient of MOS scores with TTS global cost is -0.47 (Spearman: -0.50) which confirms once again the result of section 3.2.2 for hybrid TTS.

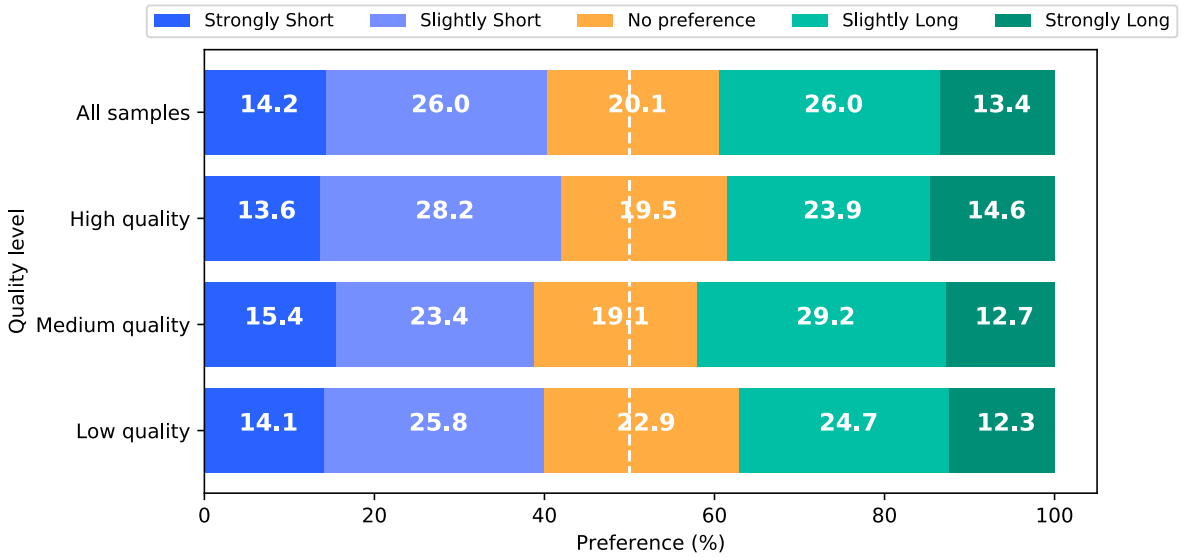
The main question of this experiment was to compare the quality of mixed signals with fully recorded speech using different levels of synthetic quality. We used voice corpus sizes to simulate the synthetic quality levels. Now we have evaluated the samples, we can sort the results based on the perceived synthetic quality which can be obtained according to MOS scores of fully synthetic signal (*SSSS* configuration). The *SSSS* samples with different voice corpus sizes are grouped to three levels of high, medium, and low quality (20 samples for each group) based on their MOS scores. In this way, mixed signals are categorized to new quality levels based on the label of their synthetic parts. For example if *SSSS* configuration of a script with 1 hour voice corpus has been labeled as high quality, all mixed configurations of this script with 1 hour voice corpus should be labeled as high quality. Indeed this new aggregation causes the script of samples in each category to be potentially different. Nevertheless, it helps to consider the samples in different perceptual quality levels.

Figure 7.3 displays the MOS test result (7.3a) and the AB test result (7.3b) based on quality levels.

Figure 7.3a shows that by improving the quality level from low to high the difference between mixed signal and fully synthetic signals decreased. The MOS score of fully synthetic signals in high synthetic quality level is comparable with mixed signals. Figure 7.3b confirms that *long* and *short* synthetic parts with different quality levels in mixed signals do not have any overall preference. Although a narrow band for *no preference* (about 20%) shows that sometimes listeners prefer *long* synthetic parts and sometimes *short* synthetic parts. It means that it was not a difficult task for listeners to tell their preference about a single sample. Anyway, their preference was not caused by the length of continuous synthetic part.



(a) MOS test result for evaluating mixed signals based on perceptual quality level



(b) AB test result for evaluation the impact of the length of continuous synthetic part in mixed signals based on perceptual quality level

Figure 7.3 – Aggregating previous perceptual tests result based on perceptual quality level of fully synthetic signals instead of voice corpus size.

7.3.2 Impact of starting and ending parts

In these perceptual tests, mixed signals comprising 4 parts with total duration of 12-18 seconds are evaluated. While each configuration in table 7.1 accompanies its inverse configuration, there is a hypothesis that starting part or ending part could bias the listeners assessment. In order to investigate this hypothesis, the MOS scores of mixed signals are aggregated into four groups: the configurations that start with natural/synthetic part and the configurations that end with natural/synthetic part. The MOS score related to these groups for different voice corpus sizes are shown in table 7.3.

	5 hours	1 hour	30 minutes
Total average	3.38±0.10	2.79±0.10	2.43±0.12
Start with synthetic (SNSN,SNNS,SSNN)	3.30±0.21	2.60±0.22	2.28±0.25
Start with natural (NNSS,NSSN,NSNS)	3.48±0.18	2.96±0.18	2.58±0.21
End with synthetic (NNSS,SNNS,NSNS)	3.48±0.18	2.90±0.20	2.70±0.23
End with natural (SNSN,NSSN,SSNN)	3.30±0.21	2.66±0.20	2.15±0.23
Start with natural and end with synthetic (NSNS,NNSS)	**3.58±0.21**	**2.92±0.25**	**2.77±0.26**

Table 7.3 – Aggregating MOS test results based on starting end ending parts for mixed signals. Signals which start with natural part and end with synthetic part are evaluated with higher quality.

This result shows a trend that mixed signals which start with a natural part and end with a synthetic part have been evaluated with a higher score. This bias indicates the weakness of our protocol for evaluating the mixed synthetic and natural speech of audio books. In the final audio book all parts are in middle (except the first and the last) and starting part and ending part will have less impact on listeners' perception. In any case, it is not possible to evaluate a full audio book with listeners.

7.4 Conclusion

In this chapter, the mixing synthetic and recorded human voice for expressive audiobook generation has been investigated. A perceptual test showed that mixed signals are preferred by listeners in comparison to fully synthetic signals. This has been ob-

served in different levels of synthetic quality which was controlled by TTS voice corpus sizes. This result confirms the first assumption of the thesis according to which the TTS voice corpus should be a part of the final audio book. Hence, recording a part of audio book and using it to synthesize the rest of the same book would help to have higher overall quality in the final audio book instead of synthesizing everything.

Regardless of synthetic length in mixed audio book, the change of signal type may impact the listeners perception. Consequently, the impact of transition times in mixed signals, half synthetic and half natural, has been investigated. The MOS scores and a direct comparison in an AB test do not show that the number of transitions could change the listeners' perception and preference. The AB test has been originally designed to study the effect of the length of continuous synthetic part in mixed signal on listeners perception. Listeners do not have any preference between a long synthetic signal and those contain two short synthetic signals.

Investigation of the perceptual quality differences, between mixed signals and fully synthetic signals, also reveals that by improving the quality of synthetic signal, these two kinds of signals become comparable. As a future work, this comparison could be done with different kinds of TTS systems and bigger voice corpora which could result in higher synthetic quality.

Our analyses on results show that listeners have a bias on starting and ending parts of 4 breath groups mixed signals. It reveals that listeners have preferred mixed signals which start with a natural part and end with a synthetic part. This result emphasizes that evaluating a longer part of mixed signals is needed. Due to perceptual test duration, longer signals limit the number of evaluations per listener and need more listeners.

CONCLUSION

The main aim of this thesis is voice corpus design for unit selection-based TTS in audio-book generation task. Chapter 2 described audio-book generation problem and previous works on voice corpus design. In our audio-book generation problem, the final audio-book is a mix of synthetic and recorded speech signals. The synthetic part is the output of a TTS system built upon the recorded part. The selection of the recorded part as the TTS voice corpus is the main concern in this thesis. Beside the optimization problem, the expressiveness of audio-books is an additional challenge.

As a first step a posterior strategy was followed in chapter 3 to find attributes of the best subset solution resulting from a voice corpus reduction approach. The computational time problem did not allow to find a subset solution in reasonable time. In chapter 4, a phoneme embedding model is proposed to do the linguistic features extraction at the utterance level. By using this embedding model and set covering-based or distribution-based method, a voice subcorpus is extracted that helps TTS to synthesize signal better that previous voice corpus design methods. A statistical analysis of best voice subcorpora leads to the idea of selecting short utterances for voice corpus. Extracting shortest utterances as voice corpus performed better than previous methods in terms of perceptual synthetic quality. Finally in chapter 7 the idea of mixing synthetic and recorded signal for final audio-book is investigated. It showed that mixed signals are preferred by listeners compared to fully synthetic signal.

In this thesis at least twelve listening tests are run to confirm the results perceptually. The main achievements of thesis experiments and the future works are follow.

Contributions

The first need for any TTS voice corpus design result is an automatic evaluation. Although perceptual tests are necessary for any conclusion, an objective measure which is able to approximate the synthetic quality can reduce the evaluation cost. By comparing several objective measures, such as TTS cost, PESQ, and DTW on MGC or MFCC features, the TTS global cost has been found with stronger correlation with perceptual

scores of synthetic quality. This comparison has been reported in section 3.2.2 for expert knowledge based TTS. A considerable correlation of TTS global cost in hybrid TTS is also observed in section 7.3.1.

In the audio-book generation problem, as we mix synthetic and recorded speech signals, the length of the recorded part determines the cost of audio-book generation. Moreover, we can observed that by increasing the recorded part as TTS voice corpus the synthetic quality should increase. The exploration of a trade-off between the length of the recorded part and the TTS synthetic quality reveals that the voice corpus size, after a threshold (1 hour), is big enough and by increasing the voice corpus size, the improvement of synthetic quality can not be distinguished perceptually. This finding enables to focus future work to TTS corpus design for less than 1 hour of voice corpus.

The main contribution of the thesis is the end-to-end method for script selection in chapter 4. A fully unsupervised method which can take into account the contextual information and transforms the discrete linguistic information into a continuous embedding space is presented. The CNN embedding model, based on linguistic information, provides an embedding vector for utterances. This embedding model proves to be efficient for script selection in the voice corpus design problem.

The proposed unsupervised subset selection method has not been designed and adapted to TTS voice corpus design and can be used in other contexts. In any time series data for which contextual information has an influential role such as database design and information extraction, the proposed method can be helpful. As a different example for usage of the proposed embedding model, it is employed for calculating the target cost in a hybrid TTS. The synthetic quality showed that it performs better than classical expert knowledge based unit selection TTS. Despite of that, an acoustic embedding model performs better than this linguistic embedding model in the hybrid TTS system. A successful model or methodology has not been found to employ acoustic information besides linguistic information for voice corpus design.

The investigation of several voice subcorpora, resulting in higher synthetic quality, shows a trend to selecting short utterances. We then proposed to simply design the voice corpus by selecting the shortest utterances first. An experiment on two audio-books, with different average length of utterances and two unit selection based TTS systems, confirmed that short utterances are more efficient than all previous methods for designing the voice corpus.

Finally, after these findings, we came back to the original problem. The initial audio-

book generation problem was based on an assumption that the overall quality achieved by mixing a recorded part of book and a synthetic part is higher than the result of a fully synthetic audio-book. Some previous works observed that fully synthetic signals are preferred rather than a mix of synthetic and natural signals in terms of naturalness by listeners. A comparison between fully synthetic and recorded signals in terms of overall quality showed that listeners preferred the mixed signal in our case. We have also found that the number of transitions between synthetic and recorded signals does not impact on listeners' preference. It means that in the voice corpus design process, the order of selected or not selected utterances for recording part is not important.

Perspectives

This thesis is centered on voice corpus design for unit selection TTS systems, however today the end-to-end TTS systems are widely developed and used. It is needed to confirm the results on state of the art vocoder-based TTS systems as well.

Although the TTS cost has been found highly correlated with perceptual evaluation, it does not take into account the prosody and intonation of the synthetic signal. It concentrates more on smoothness of synthetic signal. Another objective measure or a methodology for quality evaluation is necessary to consider other aspects of speech quality. Specially when the naturalness and smoothness of synthetic signals are good enough, the prosody of signal becomes more important.

It has been found out that the previous measures for evaluating or even designing subcorpus such as KLD and coverage rate of linguistics labels are not good enough (see section 6.4.1). An objective metric for automatic assessment of voice corpus richness could help the process of unsupervised designing. Beside linguistic information, the corresponding estimated acoustic information should be taken into account for this matter. For example using and acoustic model could help to predict the duration and the acoustic representation of phones. The coverage of acoustic diversity should be taken into account in addition to coverage of linguistic diversity for TTS voice corpus design.

The idea of using acoustic information for voice corpus design has failed (see section 5.1). While the acoustic information plays an important role in TTS context, an investigation for the reason of the failure or proposing a methodology to take into account the estimated acoustic information related to context of script besides linguistic

information is expected.

The main conclusion of this thesis is that short utterances are the best candidates for voice corpus design. The acoustic and linguistic characteristic of short utterances could be the future subject of study. The performance of selecting short utterances in other context, specially when the task is not expressive like reading news, can be studied.

The cost of audio-book generation will be reduced by synthesising a portion of the book using TTS systems. The presented approach in this study could help to have a compromise between the budget and the quality of final audio-book. But the contribution of this study is not limited to audio-book generation task. The volume of voice corpus for training TTS, ASR, speech emotion recognition, speaker recognition, etc is daily growing. The new speech corpus should be prepared based on the context and the needs of task. The voice corpus design in audio-book generation problem has raised the question of *how does the speech corpus can be optimized for an specific (or even dynamic) context?* An optimized procedure could help to save time and budget.

In a wider view, in this thesis, the problem of subset selection of sequential data has been investigated. While the objective of a subset selection is depending on the task, the optimization problem can be addressed by the same protocol. Presented model and methodologies in this thesis can be applied to other fields of study such as corpus design for automatic speech recognition, database design for machine translation, or summarization.

BIBLIOGRAPHY

Akuzawa, K., Iwasawa, Y., & Matsuo, Y., (2018), Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder, *in Nineteen Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Hyderabad, India.

Alain, P., Barbot, N., Chevelu, J., Lecorvé, G., Simon, C., & Tahon, M., (2017), The IRISA Text-To-Speech System for the Blizzard Challenge 2017, *in Blizzard Challenge workshop. International Speech Communication Association (ISCA)*, Stockholm, Sweden.

Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Et al., (2017), Deep voice: Real-time neural text-to-speech, *in Thirty-fourth International Conference on Machine Learning*, JMLR.org.

Avanzi, M., Christodoulides, G., Lolive, D., Delais-Roussarie, E., & Barbot, N., (2014), Towards the adaptation of prosodic models for expressive text-to-speech synthesis., *in Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore.

Baljekar, P., & Black, A. W., (2016), Utterance Selection Techniques for TTS Systems Using Found Speech, *in Ninth ISCA Workshop on Speech Synthesis (SSW9)*, Sunnyvale, USA.

Barbot, N., Boëffard, O., Chevelu, J., & Delhay, A., (2015), Large linguistic corpus reduction with SCP algorithms, *Computational Linguistics*, *413*, 355–383.

Black, A. W., & Campbell, N., (1995), Optimising selection of units from speech databases for concatenative synthesis., *in Fourth European Conference on Speech Communication and Technology (EUROSPEECH)*, Madrid, Spain.

Black, A. W., & Lenzo, K. A., (2001), Optimal data selection for unit selection synthesis, *in Fourth ISCA ITRW on Speech Synthesis (SSW4)*, Perthshire, Scotland.

Black, A. W., & Taylor, P. A., (1997), Automatically clustering similar units for unit selection in speech synthesis., *in Fifth European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece.

Boeffard, O., Charonnat, L., Le Maguer, S., Lolive, D., & Vidal, G., (2012), Towards Fully Automatic Annotation of Audio Books for TTS., *in Eighth International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Bonafonte, A., Adell, J., Esquerra, I., Gallego, S., Moreno, A., & Pérez, J., (2008), Corpus and Voices for Catalan Speech Synthesis, *in Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Bozkurt, B., Ozturk, O., & Dutoit, T., (2003), Text Design for TTS Speech Corpus Building Using a Modified Greedy Selection, *in Eighth European Conference on Speech Communication and Technology (EUROSPEECH)*.

Braunschweiler, N., & Buchholz, S., (2011), Automatic Sentence Selection from Speech Corpora Including Diverse Speech for Improved HMM-TTS Synthesis Quality., *in Twelfth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Florence, Italy.

Braunschweiler, N., Gales, M. J., & Buchholz, S., (2010), Lightly supervised recognition for automatic alignment of large coherent speech recordings, *in Eleventh Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japan.

Cadic, D., Boidin, C., & d'Alessandro, C., (2010), Towards Optimal TTS Corpora., *in Seventh Ninth International Conference on Language Resources and Evaluation (LREC)*, Valetta, Malta.

Campbell, N., (2003), Towards synthesising expressive speech; designing and collecting expressive speech data., *in Eighth European Conference on Speech Communication and Technology (EUROSPEECH)*, Geneva, Switzerland.

Campbell, N., (2007), Evaluation of speech synthesis, *in Evaluation of text and speech systems*, Springer.

Campbell, N., (2008), Expressive/affective speech synthesis, *in Springer Handbook of Speech Processing*, Springer.

Chalamandaris, A., Tsiakoulis, P., Karabetsos, S., & Raptis, S., (2014), Using Audio Books for Training a Text-to-Speech System, *in Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland.

Charfuelan, M., & Schröder, M., (2012), Correlation analysis of sentiment analysis scores and acoustic features in audiobook narratives, *in Fourth International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (ES3)*, Istanbul, Turkey, Citeseer.

Charfuelan, M., & Steiner, I., (2013), Expressive speech synthesis in MARY TTS using audiobook data and emotionML., *in Fourteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Lyon, France.

Chevelu, J., & Lolive, D., (2015), Do not build your TTS training corpus randomly, *in Twenty-third European Signal Processing Conference (EUSIPCO)*, Nice, France, IEEE.

Chevelu, J., Lolive, D., Maguer, S. L., & Guennec, D., (2015), How to compare TTS systems: a new subjective evaluation methodology focused on differences, *in Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany.

Chiaráin, N. N., & Chasaide, A. N., (2017), Effects of Educational Context on Learners' Ratings of a Synthetic Voice., *in Seventh ISCA Workshop on Speech and Language Technology in Education*, Stockholm, Sweden.

Chu, M., & Peng, H., (2001), An objective measure for estimating MOS of synthesized speech, *in Seventh European Conference on Speech Communication and Technology (EUROSPEECH)*, Aalborg, Denmark.

Clark, R., Silen, H., Kenter, T., & Leith, R., (2019), Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs, *in Tenth ISCA Workshop on Speech Synthesis (SSW10)*, Vienna, Austria.

Cooper, E., Chang, A., Levitan, Y., & Hirschberg, J., (2016), Data Selection and Adaptation for Naturalness in HMM-Based Speech Synthesis., *in Seventeenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, San Francisco, USA.

Do, C.-T., Evrard, M., Leman, A., d'Alessandro, C., Rilliard, A., & Crebouw, J.-L., (2014), Objective evaluation of HMM-based speech synthesis system using Kullback-Leibler divergence, *in Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore.

Dubey, R. K., & Kumar, A., (2015), Non-intrusive speech quality assessment using multi-resolution auditory model features for degraded narrowband speech, *IET Signal Processing*, *99*, 638–646.

Edge, R. L., (2012), *Measuring Speech Naturalness of Children who Do and Do Not Stutter: The Effect of Training and Speaker Group on Speech Naturalness Ratings and Agreement Scores when Measured by Inexperienced Listeners* (Doctoral dissertation), University of Georgia.

Espinosa, D., White, M., Fosler-Lussier, E., & Brew, C., (2010), Machine Learning for Text Selection with Expressive Unit-Selection Voices, *in Eleventh Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japan.

Eyben, F., Buchholz, S., & Braunschweiler, N., (2012), Unsupervised clustering of emotion and voice styles for expressive TTS, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, IEEE.

Falk, T. H., & Chan, W.-Y., (2006), Single-ended speech quality measurement using machine learning methods, *IEEE Transactions on Audio, Speech, and Language Processing*, *146*, 1935–1947.

Fan, Y., Qian, Y., Xie, F.-L., & Soong, F. K., (2014), TTS synthesis with bidirectional LSTM based recurrent neural networks, *in Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore.

Fernandez, R., Rendel, A., Ramabhadran, B., & Hoory, R., (2015), Using deep bidirectional recurrent neural networks for prosodic-target prediction in a unit-selection text-to-speech system, *in Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany.

François, H., & Boëffard, O., (2001), Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem, *in Seventh Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Aalborg, Denmark.

François, H., & Boëffard, O., (2002), The Greedy Algorithm and its Application to the Construction of a Continuous Speech Database., *in Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain.

Freedman, D., Pisani, R., & Purves, R., (2007), Statistics (international student edition), *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

Fu, S.-W., Tsao, Y., Hwang, H.-T., & Wang, H.-M., (2018), Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model based on BLSTM, *in Nineteen Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Hyderabad, India.

Fujisaki, H., (2004), Information, prosody, and modeling-with emphasis on tonal features of speech, *in International Conference of Speech Prosody*.

Gauvain, J.-L., Lamel, L., & Eskénazi, M., (1990), Design considerations and text selection for BREF, a large French read-speech corpus., *in First International Conference on Spoken Language Processing (ICSLP)*, Orsay, France.

Gong, L., & Lai, J., (2003), To mix or not to mix synthetic speech and human speech? Contrasting impact on judge-rated task performance versus self-rated performance and attitudinal responses, *International Journal of Speech Technology*, *62*, 123–131.

Grancharov, V., & Kleijn, W. B., (2008), Speech quality assessment, *in Springer Handbook of Speech Processing*, Springer.

Guo, Y., & Zhu, J., (2020), Naturalness evaluation of synthetic speech based on residual learning networks, *in Seventh Conference on Sound and Music Technology (CSMT)*, Springer.

Gupta, R., Banville, H. J., & Falk, T. H., (2017), Multimodal physiological quality-of-experience assessment of text-to-speech systems, *IEEE Journal of Selected Topics in Signal Processing*, *111*, 22–36.

Hakami, M., & Kleijn, W. B., (2017), Machine learning based non-intrusive quality estimation with an augmented feature set, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, IEEE.

Halabi, N., & Wald, M., (2016), Phonetic inventory for an Arabic speech corpus, *in Tenth International Conference on Language Resources and Evaluation (LREC)*, Paris, France.

Hayashi, T., Yamamoto, R., Inoue, K., Yoshimura, T., Watanabe, S., Toda, T., Takeda, K., Zhang, Y., & Tan, X., (2020), ESPnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE.

Hinterleitner, F., (2017), *Quality of Synthetic Speech;Perceptual Dimensions, Influencing Factors, and Instrumental Assessment*, Springer.

Hinterleitner, F., Neitzel, G., Möller, S., & Norrenbrock, C., (2011), An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks, *in Blizzard Challenge workshop. International Speech Communication Association (ISCA)*, Florence, Italy.

Holub, J., Avetisyan, H., & Isabelle, S., (2017), Subjective speech quality measurement repeatability: comparison of laboratory test results, *International Journal of Speech Technology*, *201*, 69–74.

Hu, Y., & Loizou, P. C., (2008), Evaluation of objective quality measures for speech enhancement, *IEEE Transactions on audio, speech, and language processing*, *161*, 229–238.

Hunt, A. J., & Black, A. W., (1996), Unit selection in a concatenative speech synthesis system using a large speech database, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE.

Iida, A., Campbell, N., Higuchi, F., & Yasumura, M., (2003), A corpus-based speech synthesis system with emotion, *Speech Communication*, *401*, 161–187.

Isogai, M., & Mizuno, H., (2010), Speech database reduction method for corpus-based TTS system, *in Eleventh Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japan.

Itoh, N., Sainath, T. N., Jiang, D. N., Zhou, J., & Ramabhadran, B., (2012), N-best entropy based data selection for acoustic modeling, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, IEEE.

Jauk, I., (2017), Unsupervised learning for expressive speech synthesis, *in Eighteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden.

Karp, R. M., (1972), Reducibility among combinatorial problems, *in Complexity of computer computations, The IBM Research Symposia Series*, Springer.

Kawai, H., Toda, T., Ni, J., Tsuzaki, M., & Tokuda, K., (2004), XIMERA: A new TTS from ATR based on corpus-based technologies, *in Fifth ISCA Workshop on Speech Synthesis (SSW5)*, Pittsburgh, USA.

Kawai, H., Yamamoto, S., Higuchi, N., & Shimizu, T., (2000), A design method of speech corpus for text-to-speech synthesis taking account of prosody, *in Sixth International Conference on Spoken Language Processing*.

Kendall, M. G., (1948), *Rank correlation methods.*, London, Griffin.

Kim, D.-S., (2005), ANIQUE: An auditory model for single-ended speech quality estimation, *IEEE Transactions on Speech and Audio Processing*, *135*, 821–831.

King, S., Crumlish, J., Martin, A., & Wihlborg, L., (2018), The Blizzard Challenge 2018, *in Blizzard Challenge workshop. International Speech Communication Association (ISCA)*, Hyderabad, India.

King, S., Lovisa, W., & Wei, M., (2017), The Blizzard Challenge 2017, *in Blizzard Challenge workshop. International Speech Communication Association (ISCA)*, Stockholm, Sweden.

Kominek, J., & Black, A. W., (2003), *CMU Arctic databases for speech synthesis* (tech. rep. CMU-LTI-03-177), Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA, http://www.festvox.org/cmu-arctic

Kondo, K., (2012), *Subjective Quality Measurement of Speech, Signals and Communication Technology*, Springer-Verlag Berlin Heidelberg.

Koriyama, T., & Kobayashi, T., (2015), A comparison of speech synthesis systems based on GPR, HMM, and DNN with a small amount of training data, *in Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany.

Krul, A., Damnati, G., Yvon, F., Boidin, C., & Moudenc, T., (2007), Approaches for adaptive database reduction for text-to-speech synthesis., *in Eighth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Antwerp, Belgium.

Krul, A., Damnati, G., Yvon, F., & Moudenc, T., (2006), Corpus design based on the kullback-leibler divergence for text-to-speech synthesis application, *in Ninth International Conference on Spoken Language Processing*.

Lambert, T., Braunschweiler, N., & Buchholz, S., (2007), How (not) to select your voice corpus: random selection vs. phonologically balanced., *in Sixth ISCA Workshop on Speech Synthesis (SSW6)*, Bonn, Germany.

Larnel, L. F., Gauvain, J.-L., & Eskenazi, M., (1991), BREF, a large vocabulary spoken corpus for French, *in Second European Conference on Speech Communication and Technology (EUROSPEECH)*, Genova, Italy.

Latorre, J., Yanagisawa, K., Wan, V., Kolluru, B., & Gales, M. J., (2014), Speech intonation for TTS: Study on evaluation methodology, *in Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore.

Le Maguer, S., Barbot, N., & Boëffard, O., (2013), Evaluation of contextual descriptors for HMM-based speech synthesis in French., *in Eighth ISCA Workshop on Speech Synthesis (SSW8)*, Barcelona, Spain.

Le, Q., & Mikolov, T., (2014), Distributed representations of sentences and documents, *in International Conference on Machine Learning*.

Lecun, Y., & Bengio, Y., (1995), Convolutional networks for images, speech, and time-series, *in The handbook of brain theory and neural networks*, MIT Press.

Lewis, J. R., Commarford, P. M., & Kotan, C., (2006), Web-based comparison of two styles of auditory presentation: All tts versus rapidly mixed tts and recordings, *in Human Factors and Ergonomics Society Annual Meeting*, San Francisco.

Li, Q., Fang, Y., Lin, W., & Thalmann, D., (2014), Non-intrusive quality assessment for enhanced speech signals based on spectro-temporal features, *in IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, Chengdu, China, IEEE.

Loizou, P. C., (2011), Speech quality assessment, *in Multimedia analysis, processing and communications*, Springer.

Luong, H.-T., Takaki, S., Henter, G. E., & Yamagishi, J., (2017), Adapting and controlling DNN-based speech synthesis using input codes, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, IEEE.

Merritt, T., Clark, R. A., Wu, Z., Yamagishi, J., & King, S., (2016), Deep neural network-guided unit selection synthesis, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, IEEE.

Morise, M., Yokomori, F., & Ozawa, K., (2016), WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE TRANSACTIONS on Information and Systems*, *997*, 1877–1884.

Mueller, J., & Thyagarajan, A., (2016), Siamese recurrent architectures for learning sentence similarity, *in Thirtieth AAAI conference on Artificial Intelligence*, Arizona, USA.

Ni, J., Hirai, T., & Kawai, H., (2006), Constructing a phonetic-rich speech corpus while controlling time-dependent voice quality variability for English speech synthesis, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE.

Norrenbrock, C. R., Hinterleitner, F., Heute, U., & Möller, S., (2012), Towards perceptual quality modeling of synthesized audiobooks-Blizzard Challenge 2012, *in Blizzard Challenge workshop. International Speech Communication Association (ISCA)*, Portland, USA.

Norrenbrock, C. R., Hinterleitner, F., Heute, U., & Möller, S., (2015), Quality prediction of synthesized speech based on perceptual quality dimensions, *Speech Communication*, *66*, 17–35.

Nose, T., Arao, Y., Kobayashi, T., Sugiura, K., & Shiga, Y., (2017), Sentence Selection Based on Extended Entropy Using Phonetic and Prosodic Contexts for Statis-

tical Parametric Speech Synthesis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *255*, 1107–1116.

Nose, T., Arao, Y., Kobayashi, T., Sugiura, K., Shiga, Y., & Ito, A., (2015a), Entropy-based sentence selection for speech synthesis using phonetic and prosodic contexts, *in Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany.

Nose, T., Arao, Y., Kobayashi, T., Sugiura, K., Shiga, Y., & Ito, A., (2015b), Entropy-based sentence selection for speech synthesis using phonetic and prosodic contexts, *in Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K., (2016), Wavenet: A generative model for raw audio, *arXiv preprint arXiv:1609.03499*.

P.862, I.-T. R., (2001), Perceptual Evaluation Of Speech Quality (PESQ), An Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs, *International Telecommunication Union*.

Perquin, A., Lecorvé, G., Lolive, D., & Amsaleg, L., (2018), Phone-level embeddings for unit selection speech synthesis (P. G. Dutoit T. Martín-Vide C., Ed.), *in* P. G. Dutoit T. Martín-Vide C. (Ed.), *International Conference on Statistical Language and Speech Processing, LNCS/LNAI*, Springer.

Prahallad, K., & Black, A. W., (2010), Segmentation of monologues in audio books for building synthetic voices, *IEEE Transactions on Audio, Speech, and Language Processing*, *195*, 1444–1449.

Recommendation, I., (2003), 1534-1: Method for the subjective assessment of intermediate quality level of coding systems, *International Telecommunication Union*.

Recommendation, I., (1996), 800, Methods for subjective determination of transmission quality, *International Telecommunication Union*.

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P., (2001), Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, USA, IEEE.

Rothauser, E., (1969), IEEE recommended practice for speech quality measurements, *IEEE Transactions on Audio and Electroacoustics*, *17*, 225–246.

Rouibia, S., & Rosec, O., (2005), Unit selection for speech synthesis based on a new acoustic target cost, *in Ninth European Conference on Speech Communication and Technology (EUROSPEECH)*, Lisbon, Purtugal.

Sagisaka, Y., Kaiki, N., Iwahashi, N., & Mimura, K., (1992), ATR - Talk Speech Synthesis System, *in Second International Conference on Spoken Language Processing*, Alberta, Canada.

Saito, Y., Takamichi, S., & Saruwatari, H., (2017), Statistical parametric speech synthesis incorporating generative adversarial networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *261*, 84–96.

Saratxaga, I., Navas, E., Hernáez, I., & Luengo, I., (2006), Designing and Recording an Emotional Speech Database for Corpus Based Synthesis in Basque., *in Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Shamsi, M., (2020), TTS voice corpus reduction for audio-book generation, *in 6^{e} conférence conjointe Journées d'Études sur la Parole (JEP, 31^{e} édition), Traitement Automatique des Langues Naturelles (TALN, 27^{e} édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22^{e} édition)*, Nancy, France.

Shamsi, M., Barbot, N., Lolive, D., & Chevelu, J., (2020), Mixing synthetic and recorded signals foraudio-book generation, *in Twenty-second International Conference on Speech and Computer (SPECOM)*, St. Petersburg, Russia, Springer.

Shamsi, M., Chevelu, J., Lolive, D., & Barbot, N., (2020), Corpus design for expressive speech: impact of the utterance length, *in Tenth International Conference of Speech Prosody*, Tokyo, Japan.

Shamsi, M., Lolive, D., Barbot, N., & Chevelu, J., (2019a), Corpus Design using Convolutional Auto-Encoder Embeddings for Audio-Book Synthesis, *in Twentieth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria.

Shamsi, M., Lolive, D., Barbot, N., & Chevelu, J., (2019b), Investigating the relation between voice corpus design and hybrid synthesis under reduction constraint, *in Seventh International Conference on Statistical Language and Speech Processing (SLSP)*, Ljubljana, Slovenia, Springer.

Shamsi, M., Lolive, D., Barbot, N., & Chevelu, J., (2019c), Script Selection using Convolutional Auto-encoder for TTS Speech Corpus, *in Twenty-first International Conference on Speech and Computer (SPECOM)*, Istanbul, Turkey, Springer.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Et al., (2018), Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE.

Shinohara, Y., (2014), A submodular optimization approach to sentence set selection, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, IEEE.

Sini, A., Lolive, D., Vidal, G., Tahon, M., & Delais-Roussarie, E., (2018), SynPaFlex Corpus: An Expressive French Audiobooks Corpus dedicated to expressive speech synthesis, *in Eleventh International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan.

Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., & Bengio, Y., (2017), Char2wav: End-to-end speech synthesis, *in Fifth International Conference on Learning Representations (ICLR)*.

Spearman, C., (1904), The Proof and Measurement of Association between Two Things, *The American Journal of Psychology*, *151*, 72–101.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R., (2014), Dropout: a simple way to prevent neural networks from overfitting, *The Journal of Machine Learning Research*, *151*, 1929–1958.

Stanton, D., Wang, Y., & Skerry-Ryan, R., (2018), Predicting expressive speaking style from text in end-to-end speech synthesis, *in Seventh IEEE Workshop on Spoken Language Technology (SLT)*, IEEE.

Sutskever, I., Vinyals, O., & Le, Q. V., (2014), Sequence to sequence learning with neural networks, *in Advances in Neural Information Processing Systems (NIPS)*.

Székely, E., Cabral, J. P., Cahill, P., & Carson-Berndsen, J., (2011), Clustering Expressive Speech Styles in Audiobooks Using Glottal Source Parameters., *in Twelfth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Florence, Italy.

Székely, E., Csapó, T. G., Tóth, B., Mihajlik, P., & Carson-Berndsen, J., (2012), Synthesizing expressive speech from amateur audiobook recordings, *in Fourth IEEE Workshop on Spoken Language Technology (SLT)*, Miami, USA, IEEE.

Szklanny, K., & Koszuta, S., (2017), Implementation and verification of speech database for unit selection speech synthesis, *in Federated Conference on Computer Science and Information Systems (FedCSIS)*, Prague, Czech Republic, IEEE.

Tahon, M., Lecorvé, G., Lolive, D., & Qader, R., (2017), Perception of expressivity in TTS: linguistics, phonetics or prosody?, *in International Conference on Statistical Language and Speech Processing, LNCS/LNAI*, Springer.

Theune, M., Meijs, K., Heylen, D., & Ordelman, R., (2006), Generating expressive speech for storytelling applications, *IEEE Transactions on Audio, Speech, and Language Processing*, *144*, 1137–1144.

Toda, T., Black, A. W., & Tokuda, K., (2004), Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis, *in Fifth ISCA Workshop on Speech Synthesis (SSW5)*, Pittsburgh, USA.

Toda, T., Kawai, H., & Tsuzaki, M., (2004), Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Canada, IEEE.

Toda, T., Kawai, H., Tsuzaki, M., & Shikano, K., (2006), An evaluation of cost functions sensitively capturing local degradation of naturalness for segment selection in concatenative speech synthesis, *Speech Communication*, *481*, 45–56.

Ullmann, R., Rasipuram, R., Magimai.-Doss, M., & Bourlard, H., (2015), Objective intelligibility assessment of text-to-speech systems through utterance verification, *in Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany.

Umbert, M., Moreno, A., Agüero, P. D., & Bonafonte, A., (2006), Spanish Synthesis Corpora, *in Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Van Niekerk, D. R., van Heerden, C., Kleynhans, N., Kjartansson, O., Jansche, M., Ha, L., & Davel, M. H., (2017), Rapid development of TTS corpora for four South African languages, *in Eighteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden.

Van Santen, J. P., & Buchsbaum, A. L., (1997), Methods for optimal text selection., *in Fifth European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., & Polosukhin, I., (2017), Attention is All You Need, *in 31st International Conference on Neural Information Processing Systems*, NY, USA.

Wan, V., Agiomyrgiannakis, Y., Silen, H., & Vit, J., (2017), Google's Next-Generation Real-Time Unit-Selection Synthesizer Using Sequence-to-Sequence LSTM-Based Autoencoders., *in Eighteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden.

Wang, X., Takaki, S., & Yamagishi, J., (2016), A Comparative Study of the Performance of HMM, DNN, and RNN based Speech Synthesis Systems Trained on Very Large Speaker-Dependent Corpora, *in Ninth ISCA Workshop on Speech Synthesis (SSW9)*, Sunnyvale, USA.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Et al., (2017), Tacotron: Towards end-to-end speech synthesis, *in Eighteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., & Ochiai, T., (2018), ESPnet: End-to-End Speech Processing Toolkit, *in Nineteen Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Hyderabad, India.

Watts, O., Henter, G. E., Merritt, T., Wu, Z., & King, S., (2016), From HMMs to DNNs: where do the improvements come from?, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, IEEE.

Wester, M., Valentini-Botinhao, C., & Henter, G. E., (2015), Are we using enough listeners? no!-an empirically-supported critique of interspeech 2014 TTS evaluations., *in Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany.

Wester, M., Watts, O., & Henter, G. E., (2016), Evaluating comprehension of natural and synthetic conversational speech, *in Eighth International Conference of Speech Prosody*, Boston, USA.

Wu, Z., Watts, O., & King, S., (2016), Merlin: An Open Source Neural Network Speech Synthesis System., *in Ninth ISCA Workshop on Speech Synthesis (SSW9)*, Sunnyvale, USA.

Yoshimura, T., Henter, G. E., Watts, O., Wester, M., Yamagishi, J., & Tokuda, K., (2016), A Hierarchical Predictor of Synthetic Speech Naturalness Using Neural Networks., *in Seventeenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, San Francisco, USA.

Zen, H., Senior, A., & Schuster, M., (2013), Statistical parametric speech synthesis using deep neural networks, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, IEEE.

Zen, H., Tokuda, K., & Black, A. W., (2009), Statistical parametric speech synthesis, *Speech Communication*, *5111*, 1039–1064.

Zhang, T., Chen, Z., Wu, J., Lai, S., Lei, W., & Isert, C., (2016), Objective Evaluation Methods for Chinese Text-To-Speech Systems., *in Seventeenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, San Francisco, USA.

Zhang, W., Liu, Y., Deng, Y., & Pang, M., (2010), Automatic Construction for a TTS Corpus with Limited Text, *in International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, Changsha, China, IEEE.

Zhang, Y.-J., Pan, S., He, L., & Ling, Z.-H., (2019), Learning latent representations for style control and transfer in end-to-end speech synthesis, *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE.

Zhou, X., Ling, Z.-H., Zhou, Z.-P., & Dai, L.-R., (2018), Learning and Modeling Unit Embeddings for Improving HMM-based Unit Selection Speech Synthesis, *in Nineteen Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Hyderabad, India.

UNIVERSITÉ DE RENNES 1

Titre : Optimisation de script pour la conception de corpus vocaux de TTS dans la génération de livres audio

Mot clés : sélection de script, génération de livres audio expressifs, réduction de voix, évaluation de la qualité de voix de synthèse, systèmes de synthèse de parole hybride, plongements linguistiques et acoustiques

Résumé : L'objectif de cette thèse est la génération d'un livre audio expressif, vocalisé à l'aide de signaux de parole synthétiques et naturels, avec une haute qualité et un coût d'enregistrement minimal. La stratégie consiste à sélectionner une partie du livre dont les signaux enregistrés issus de sa lecture forment une voix. Cette voix est utilisée pour vocaliser le reste du livre à l'aide d'un système de synthèse de parole. Plusieurs stratégies sont proposées successivement : une stratégie a posteriori reposant sur des techniques de réduction de corpus, l'utilisation d'un auto-encodeur basé sur un réseau neuronal (CNN) se concentrant sur les informations linguistiques, et enfin la sélection des phrases les plus courtes. Ces différentes approches sont évaluées de manière objective et subjective. Enfin, la qualité d'un livre audio mixant signaux de parole naturels et synthétiques est étudiée. Les évaluations montrent que le mélange de signaux synthétiques et naturels est préférable à une vocalisation entièrement synthétique à l'aide d'un système TTS par sélection d'unités. Ce résultat est contraire à ce qui a été rapporté dans la littérature.

Title: Script optimization for TTS voice corpus design in audio-book generation

Keywords: script selection, expressive audio-book generation, voice reduction, synthetic speech quality evaluation, hybrid TTS systems, linguistic and acoustic embeddings

Abstract: The objective of this thesis is the generation of a high quality expressive audio-book, using natural and synthetic speech signals with a minimal recording cost. The strategy consists on selecting a part of the book and recording its reading to build a voice corpus. This voice is then used for synthesizing the rest of the book using a Text-to-Speech system. Several strategies are successively proposed: a posterior approach using voice reduction methods, a neural network based (CNN) auto-encoder focusing on linguistic information, and then the selection of the shortest utterances. These different approaches are objectively and perceptually evaluated. Finally, the quality of audio-book mixing natural and synthetic speech signals is evaluated. The evaluations show the mixture of synthetic and natural signals is preferred than fully synthetic signals produced by a unit selection based TTS system.