CHAPITRE V
Corpus de données

Ce chapitre décrit l'ensemble des corpus de données utilisés pour les différents tests. Ces corpus de données proviennent d'enregistrements obtenus à partir de locuteurs volontaires, coopératifs et prévenus à l'avance. Les corpus de données utilisés sont dans cette étude les chiffres isolés (0 à 9), les 36 mots isolés (ou expressions) du vocabulaire du Trégor (voir annexe 2), les nombres à deux chiffres (00 à 99) et les villes épelées (120 villes de France).

Les données sont issues de séances spéciales d'enregistrement à travers le réseau téléphonique, elles ne proviennent pas d'enregistrements obtenus sur les serveurs vocaux mis en exploitation. L'ensemble des locuteurs utilisés pour l'enregistrement des corpus provient de différentes régions de France. Si le paramètre accent est présent dans les enregistrements, la catégorie socioprofessionnelle n'est pas prise en compte.

La procédure d'enregistrement est semi automatique. L'opératrice appelle par téléphone un locuteur se trouvant dans son bureau ou chez lui. Elle lui décrit la tâche à effectuer, en prenant note des caractéristiques du locuteur (age, sexe, ...) et des conditions téléphoniques (type de téléphone, lieu d'appel, ...). Le locuteur doit répéter par la suite les mots prononcés par le système. Si le locuteur parle trop fort, le système lui demande de parler moins fort. Si rien n'est détecté, le système lui répète le mot à prononcer. La liste prononcée par chaque locuteur est présentée dans un ordre aléatoire (variant d'un locuteur à un autre).

Après la phase d'enregistrement, une opératrice écoute chaque enregistrement[1] et le marque "**Correct**" dans le cas ou la prononciation est correcte, ou "**Mauvais**" dans le cas contraire. Pour les tests, on utilise uniquement les enregistrements étiquetés "**Correct**".

---

[1] On appelle enregistrement le signal de parole correspondant à la prononciation d'un chiffre, d'un mot, d'une expression ou d'une phrase.

Le tableau suivant présente les quatre corpus utilisés, le nombre de répétitions par locuteur, le nombre de données (locuteurs et enregistrements) ayant servi pour la phase d'apprentissage des modèles et le nombre de données ayant servi pour la phase de mesure des performances de reconnaissance. Lors du découpage de chaque corpus de données en deux sous ensembles de données, apprentissage et test, on a tenu compte de la répartition géographique des locuteurs. De ce fait, il y a sensiblement la même proportion de locuteurs de chaque région en test et en apprentissage.

Corpus	Locuteurs		Enregistrements	
	Test	Apprentissage	Test	Apprentissage
Chiffres (10 mots)	388	382	3622	3555
Trégor (36 mots)	384	381	12842	12719
Villes épelées[1]	90	90	1425	1633
Nombres à 2 chiffres	397	392	7288	7304

Tableau V.1 Nombre de locuteurs et d'enregistrements par corpus de données

[1] 180 locuteurs ont épelé de façon naturelle 20 noms de villes pris dans une liste en contenant 120. Aucune consigne n'a été donnée aux locuteurs sur la manière d'épeler les noms de villes. On peut donc avoir des courtes pauses entre chaque lettre. La base de données complète contient environ 3000 épellations de noms de villes, ce qui représente environ 20000 lettres.