Mesures de confiance

Sommaire

Les performances d'un système ASR peuvent être souvent altérées par différents paramètres comme un environnement bruité, la variabilité entre locuteurs, les disfluences inhérentes à la parole spontanée, etc. Il est donc nécessaire pour un tel système de pouvoir, de manière automatique, évaluer la fiabilité des solutions données par le système. En d'autres termes, la *mesure de confiance* associée à une hypothèse h du système ASR ($MC(h)$) peut être assimilée à la probabilité que l'hypothèse soit correcte. Celle-ci doit être comprise dans l'intervalle $[0, 1]$ et, idéalement, une valeur de 0 pour la mesure de confiance correspond à une hypothèse incorrecte, et une valeur de 1 à une hypothèse correcte.
De nombreux domaines du traitement de la parole utilisent les mesures de confiance (Lee, 2001). On les retrouve par exemple dans la reconnaissance de la parole (Wessel et Ney, 2005; Cox et Dasmahapatra, 2002; Soong et al., 2004; Ketabdar et al., 2006), dans les système de dialogue (San-Segundo et al., 2001; Raymond et al., 2004; Raymond, 2005),

dans l'identification des langues (Metze et al., 2000) ou dans la reconnaissance du locuteur (Preti et al., 2007). Dans ces domaines, les mesures de confiance peuvent être appliquées à plusieurs niveaux : au niveau du phonème (principalement dans la reconnaissance de la parole), du mot, de la phrase, des concepts (unité sémantique permettant d'exprimer le sens d'une séquence de un ou plusieurs mots de façon conceptuelle, principalement utilisée dans les systèmes de dialogue (Kobus, 2006)) ou bien au niveau de la phrase. Dans la suite de ce chapitre on se placera au niveau du mot en ce qui concerne l'utilisation des mesures de confiance, sachant que l'utilisation des mesures de confiance dans le cadre des travaux présentés ultérieurement dans cette thèse se fait également au niveau du mot.
L'article (Jiang, 2005) propose une classification des mesures de confiance en trois catégories distinctes :

1. Une grande majorité des travaux utilise les paramètres prédictifs dans le calcul des mesures de confiance. Un paramètre peut être appelé paramètre prédictif si la distribution de probabilité des mots reconnus comme étant corrects est différente de la distribution de probabilité des mots incorrects. Ces paramètres sont généralement collectés pendant le décodage et sont ensuite combinés afin d'obtenir une seule mesure indiquant le degré de véracité du mot. Ils incluent des paramètres acoustiques ainsi que des paramètres provenant du modèle de langage ou du comportement de l'algorithme de recherche.
2. La probabilité *a posteriori* d'un mot est souvent utilisée en tant que mesure de confiance étant donné le fait qu'elle représente une mesure absolue de la fiabilité d'une décision. La probabilité *a posteriori* est une estimation de la vraisemblance entre le mot w et la suite de vecteurs d'observations acoustiques X. Comme nous le verrons par la suite elle est assez difficile à calculer (voir 2.3), d'où les différentes méthodes proposées afin d'obtenir la meilleure approximation possible. Ces méthodes sont des méthodes simples, utilisant des modèles de type *filler*, aux approches plus complexes basées sur les graphes de mots.
3. Si les deux premières catégories présentent des mesures de confiance au niveau mot, de nombreux travaux ont été menés sur l'utilisation des mesures de confiance au niveau énoncé afin de vérifier le contenu d'une hypothèse (*utterance verification*). Une fois le décodage effectué, le système produit une hypothèse W dont on évalue la fiabilité à travers une mesure de confiance. L'estimation de la mesure de confiance est formulée ici comme un test statistique pour vérifier si l'hypothèse est correcte ou incorrecte.

Les travaux de cette thèse se concentrent sur l'utilisation des mesures de confiance au niveau mot et plus précisément sur l'utilisation de la probabilité *a posteriori* du mot comme mesure de confiance. Dans ce chapitre nous présentons, tout d'abord, différentes techniques d'évaluation des mesures de confiance dans la section 2.1. Nous détaillons ensuite les mesures de confiance au niveau mot basées sur les paramètres prédictifs dans la section 2.2 et sur la probabilité *a posteriori* dans la section 2.3. La dernière partie 2.4, décrit l'algorithme *Forward-Backward* que nous avons adapté pour le calcul des probabilités *a posteriori* sur les graphes de mots. Une nouvelle méthode de normalisation des variables de l'algorithme *Forward-Backward* est aussi décrite. Étant donné

que les mesures de confiance au niveau de la phrase ne font pas l'objet des travaux de cette thèse, nous ne détaillons pas ce point.

2.1 Evaluation des mesures de confiance

Il existe différentes métriques (Siu et Gish, 1999) pour évaluer les mesures de confiance. Quelques unes des plus simples sont détaillées ici, avec la précision qu'une partie de ces métriques, comme la Detection Error Tradeoff ou l'entropie croisée, seront également utilisées pour l'évaluation des travaux effectués.

2.1.1 Detection Error Tradeoff

La courbe *Detection Error Tradeoff* (DET) permet d'évaluer la capacité d'une mesure de confiance à accepter ou rejeter une hypothèse en faisant varier la valeur du seuil fixé pour cette mesure. En faisant varier la valeur du seuil α sur la mesure de confiance, la décision est prise de la manière suivante :

$$\text{hypothèse} = \begin{cases} \text{acceptée} & \text{si MC(hypothèse)} \geq \alpha \\ \text{rejetée} & \text{sinon} \end{cases} \tag{2.1}$$

L'équation 2.1 peut donc conduire à deux types d'erreurs :

- Une erreur de fausse acceptation. Appelée aussi *Fausse Alarme (FA)*, cette erreur survient dans le cas où une hypothèse est acceptée comme étant correcte alors qu'elle est incorrecte.
- Une erreur de rejet à tort. Appelée aussi *Faux Rejet (FR)*, cette erreur survient dans le cas où une hypothèse est considérée comme incorrecte alors qu'elle est correcte.

Au vu de ces deux types d'erreurs, il est possible de calculer deux taux sur un corpus d'évaluation :

- Le taux de fausse alarme :

$$FA = \frac{\text{Nombre d'hypothèses acceptées à tort}}{\text{Nombre total d'hypothèses}} \tag{2.2}$$

- Le taux de faux rejet :

$$FR = \frac{\text{Nombre d'hypothèses rejetées à tort}}{\text{Nombre total d'hypothèses}} \tag{2.3}$$

Si on considère les distributions des mesures de confiance calculées sur l'ensemble des hypothèses correctes et des hypothèses incorrectes dans la figure 2.1 on peut visualiser les taux de faux rejet (FR) et de fausses alarmes (FA) en fonction du seuil α. Quand le seuil se déplace vers la droite (sa valeur augmente), le taux de FA diminue alors que le taux de FR augmente. Si le seuil se déplace vers la gauche (sa valeur diminue), les tendances sont inversées, avec un taux de FA qui augmente et un taux de FR qui diminue. Pour chaque valeur du seuil α, un couple (FA, FR) peut être calculé. Ce couple de valeurs détermine ce qu'on appelle un point de fonctionnement du système.

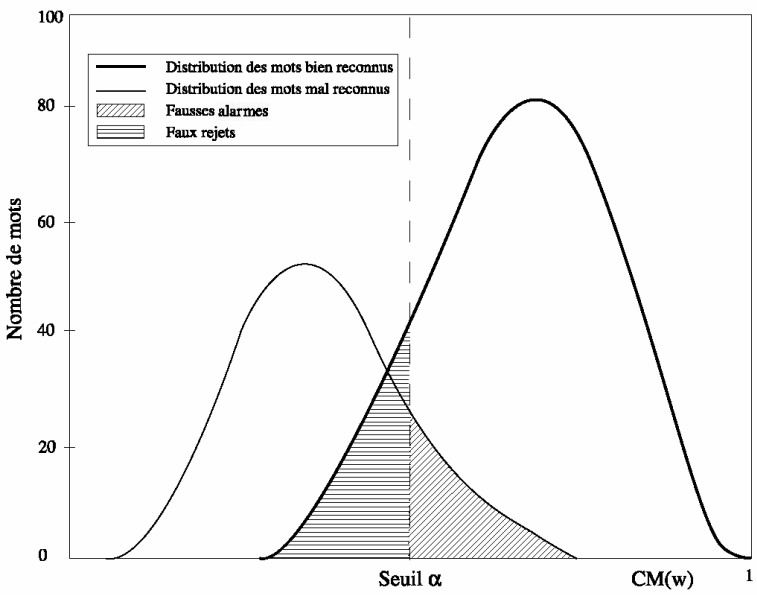

FIGURE 2.1 – Distribution des mesures de confiance sur les hypothèses de reconnaissance

La courbe DET, voir figure 2.2, décrit le taux de faux rejets (FR) en fonction du taux de fausses alarmes (FA) et permet de visualiser les différents points de fonctionnement du système qui correspondent à une valeur du seuil α variant de 0 à 1. Cette courbe, dans sa globalité, est plus adaptée dans la comparaison des performances de plusieurs systèmes que la comparaison d'un seul point de fonctionnement pour une valeur de seuil donnée. Néanmoins, un point de la courbe DET souvent utilisé dans la comparaison des systèmes est le point où les deux taux d'erreur sont égaux (EER : *Equal Error Rate*). Il se trouve à l'intersection de la courbe DET avec la droite d'égale erreur. Pour un système, plus ce point est proche de l'origine plus les mesures de confiance sont discriminantes.
Il existe aussi une autre variante de cette courbe, appelée courbe ROC (*Receiver Operating Characteristic* ou *Relative Operating Characteristic*) (Martin et al., 1997) qui, en général, décrit le taux de bonnes détections en fonction de taux de fausses alarmes.

2.1.2 Confidence Accuracy et Confidence Error Rate

Un autre moyen d'évaluer les mesures de confiance réside dans le calcul du taux d'erreur de confiance, le *Confidence Error Rate* (CER), et de son contraire, la *Confidence Accuracy* (CA). Ces deux métriques se calculent de la manière suivante :

$$CA = \frac{\text{Nombre d'étiquettes correctement assignées}}{\text{Nombre total d'étiquettes}} \tag{2.4}$$

$$CER = \frac{\text{Nombre d'étiquettes incorrectement assignées}}{\text{Nombre total d'étiquettes}} \tag{2.5}$$

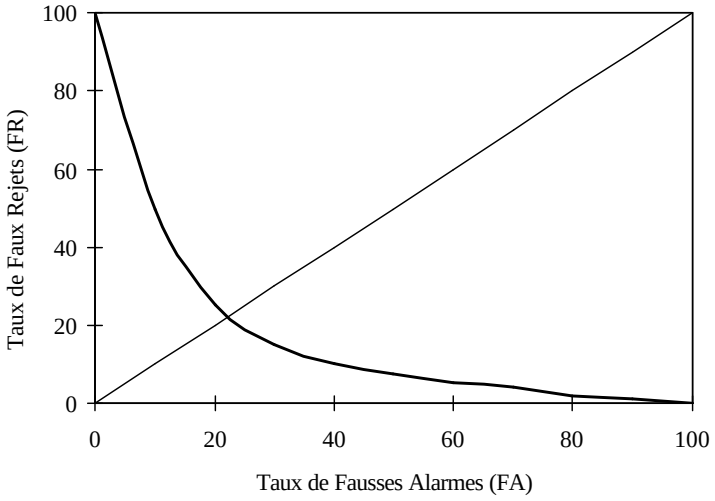

FIGURE 2.2 – Exemple de courbe DET

Le dénominateur des deux équations (*Nombre total d'étiquettes*) est le nombre total de mots reconnus par le système de reconnaissance. Chaque mot se voit désigner une étiquette, qui est en fait un attribut qui peut prendre deux valeurs, "correct" ou "incorrect". On dit qu'une étiquette est correctement assignée si à un mot reconnu comme étant correct on a assigné l'étiquette "correct" ou à un mot reconnu à tort on a assigné l'étiquette "incorrect". Le contraire signifie que l'étiquette a été incorrectement assignée. La décision d'assigner l'étiquette "correct"/"incorrect" à un mot dépend de la valeur du seuil appliqué sur la mesure de confiance du mot. Le seuil doit être optimisé sur un corpus de développement de telle façon à ce que la valeur du *Confidence Accuracy* soit la plus grande possible et inversement la valeur du *Confidence Error Rate* soit la plus petite possible. Suite à ce calibrage, la valeur du seuil obtenue peut être utilisée sur un corpus de test afin d'évaluer les performances de la mesure de confiance.

2.1.3 Entropie croisée

Les mesures de confiance doivent permettre d'estimer avec fiabilité la probabilité qu'une hypothèse soit correcte étant donné un jeu de mesures de confiance associé. Afin de choisir le jeu de mesures de confiance le plus adapté plusieurs méthodes peuvent être utilisées afin d'en évaluer leur pertinence. Une de ces méthodes consiste à mesurer la diminution relative de l'entropie croisée engendrée par la mesure de confiance.
Soit un corpus de test $\mathcal{C}$ constitué de N hypothèses de mot ; l'entropie croisée H se calcule de la manière suivante :

$$H = -\frac{1}{N}\sum_{i=1}^{N}(\delta_i \log p_i + (1-\delta_i)\log(1-p_i)) \tag{2.6}$$

δ_i est un facteur égal à 1 si l'hypothèse du mot w_i est correct et 0 sinon. La probabilité p_i représente la probabilité que l'hypothèse du mot w_i soit correcte. Sans utiliser la mesure de confiance, cette probabilité peut être évaluée comme étant la probabilité moyenne qu'un mot reconnu soit correct, ce qui est égal à la précision ($Prec$) sur le corpus de test :

$$Prec = \frac{\sum_{i=1}^{N} \delta_i}{N} \tag{2.7}$$

On peut définir ainsi l'entropie croisée initiale du corpus de test, notée H_{init} :

$$H_{init} = -Prec \log(Prec) - (1 - Prec) \log(1 - Prec) \tag{2.8}$$

Étant donné un jeu de mesures de confiance $MC = mc_1, \ldots, mc_n (n >= 1)$, la probabilité p_i devient $p_i = P(Cor|MC(w_i))$, la probabilité $P(Cor|MC(w_i))$ étant la probabilité que l'hypothèse de mot w_i soit correcte étant donné le jeu de mesures de confiance associé $MC(w_i)$. L'entropie croisée du corpus de test se calcule alors :

$$H_{MC} = -\frac{1}{N} \sum_{i=1}^{N} (\delta_i \log P(Cor|MC(w_i)) + (1 - \delta_i) \log(1 - P(Cor|MC(w_i)))) \tag{2.9}$$

Une mesure de confiance est d'autant plus efficace que l'information additionnelle apportée à l'hypothèse de mot est grande. L'évaluation se fait en calculant la diminution relative induite sur l'entropie croisée d'un corpus de test donnée, noté ΔH, par le jeu de mesures de confiance choisi :

$$\Delta H = \frac{H_{init} - H_{MC}}{H_{init}} \cdot 100 \tag{2.10}$$

Pour une mesure de confiance purement aléatoire il n'y a aucune information additionnelle et donc l'entropie du corpus de test reste inchangée. Par conséquence, la diminution relative est nulle. En revanche, pour une mesure de confiance "idéale", la prise de décision sur l'exactitude d'une hypothèse est parfaite, et donc $H_{MC} = 0$. La valeur de la diminution relative ΔH varie alors de 0 à 100%. Plus ΔH est élevée, plus l'information additionnelle apportée est importante et la mesure de confiance prédictive.

2.2 Paramètres prédictifs

Un paramètre prédictif idéal doit fournir une information suffisante afin de pouvoir départager les mots reconnus qui sont corrects de ceux reconnus à tort de telle manière que le recouvrement des distributions des deux classes soit le plus petit possible.

2.2.1 Paramètres acoustiques

Un paramètre simple évoqué dans (Jiang, 2005) est le score acoustique par trame. Celui-ci est obtenu en normalisant la vraisemblance acoustique du mot par le nombre

de trames acoustiques du mot. Toutefois, ce paramètre ne permet pas à lui seul d'obtenir une estimation précise de la pertinence du mot.

Une autre approche utilisant le score acoustique vise à isoler l'information acoustique de celle provenant du modèle de langage en utilisant, en parallèle avec le moteur de reconnaissance, une boucle de phonèmes non contrainte par un modèle de langage. Cette boucle permet à n'importe quel phonème d'en suivre un autre avec une probabilité égale et le décodage ne se base que sur la vraisemblance acoustique entre le signal et le modèle acoustique. L'utilisation d'une boucle de phonèmes n'est pas contraignante en termes de complexité de calcul. Celle-ci est très réduite par rapport à un moteur de reconnaissance intégrant un modèle de langage. Le ratio entre le score acoustique du mot reconnu et le score des phonèmes obtenus par la boucle de phonèmes sur le même intervalle temporel est proposé comme mesure de confiance dans (Young et al., 1997).

Une autre façon de calculer les mesures de confiance est proposée dans (Cox et Dasmahapatra, 2002) qui utilise des techniques de corrélation entre phonèmes. Ainsi, une mesure de confiance proposée est estimée en calculant une matrice de confusion croisée entre la transcription en phonèmes du mot reconnu, donnée par le dictionnaire, et la séquence de phonèmes reconnus par le décodeur. Les valeurs de la matrice sont estimées en réalisant un alignement entre les deux séquences et en comptant le nombre de phonèmes correctement alignés. Étant donné que cette méthode ne tient pas compte du fait que le mot ou la séquence de phonèmes soient corrects ou non, les auteurs proposent d'utiliser un corpus de développement afin de construire deux matrices de confusion croisées entre la séquence de phonèmes et les mots corrects d'un coté et les mots incorrects de l'autre coté. Ces deux matrices sont ensuite utilisées pour estimer la probabilité qu'un mot soit correct ou incorrect étant donné l'alignement des séquences de phonèmes dans les deux matrices. La mesure de confiance est estimée comme étant le rapport de vraisemblance entre ces deux probabilités.

Un autre paramètre évoqué dans (Jiang, 2005; Wessel et al., 2001) est la stabilité acoustique. La raison qui pousse au calcul de ce paramètre consiste dans le fait qu'un mot a une probabilité élevée d'être correct s'il est présent dans la même position, donnée par l'alignement de Levensthein, dans une majorité des phrases décodées par le SRAP utilisant différentes techniques ou paramètres. Généralement, il s'agit soit de faire une comparaison sur les N meilleures hypothèses issues du décodage soit de réaliser plusieurs décodages en variant différents paramètres. Dans une grande majorité des cas on effectue plusieurs décodages en variant la valeur du *fudge* afin d'obtenir la meilleure hypothèse du SRAP pour chaque décodage. La stabilité acoustique est définie comme le rapport entre le nombre de fois où le mot apparaît dans les hypothèses alternatives et le nombre total d'hypothèses. Dans (Wessel et al., 2001) la stabilité acoustique obtient de bonnes performances sur la majorité de corpus utilisés (ARISE, Verbmobil, Broadcast News, NAB). De très bonnes performances sont obtenues sur le corpus ARISE, qui peuvent s'expliquer par une longueur moyenne des énoncés très faible. Toutefois, la différence entre les performances de la stabilité acoustique et celles de la probabilité *a posteriori* calculée sur les graphes de mots (voir 2.3) est assez significative.

Discussions

Le but principal de ces techniques utilisées pour le calcul des mesures de confiance est d'essayer d'isoler le modèle de langage et le modèle acoustique afin de pouvoir calculer des mesures de confiance basées uniquement sur la modélisation acoustique des mots. En effet, dans (Palmer et Ostendorf, 2001) et (Cox et Dasmahapatra, 2002) on a observé que la probabilité conditionnelle qu'un mot soit correct ou incorrect dépend du fait que le mot précédent est correct ou incorrect. Ceci suggère un découplage entre les deux modèles afin d'éliminer les effets corrélatifs du modèle de langage. Toutefois, un découplage total des deux modèles n'est pas possible. Les deux approches (Young et al., 1997; Cox et Dasmahapatra, 2002), présentées ci-dessus, proposent d'utiliser une boucle de phonèmes en parallèle avec le SRAP. Les performances des mesures de confiance observées dans (Cox et Dasmahapatra, 2002) sur un corpus extrait du *Wall Street Journal database* ne sont pas très convaincantes et sont très loin des performances obtenues avec la stabilité acoustique.

2.2.2 Paramètres linguistiques

De la même manière que le modèle acoustique, le modèle de langage peut lui aussi constituer une source dans le calcul des paramètres prédictifs. On peut par exemple utiliser des paramètres comme le score du modèle de langage ou le comportement du repli (le degré du modèle de repli [1]) directement comme mesure de confiance(San-Segundo et al., 2001). La technique du repli consiste à utiliser un modèle de langage plus général lorsqu'un modèle spécifique ne détient pas les informations nécessaires. Plus précisément, le système peut faire appel à un $(n-1)$-gramme si une séquence des n mots n'a pas été vue dans le corpus d'apprentissage pour un n-gramme. Une mesure de confiance est ainsi attribuée en fonction du degré du repli. Dans (Uhrik et Ward, 1997), les auteurs proposent d'attribuer un score de confiance arbitraire au mot considéré, en fonction du degré de repli pour les deux mots précédents ou encore pour le mot précédent et le mot suivant.

Un autre type de paramètre est celui lié au *parsing* des hypothèses de reconnaissance. Par exemple, dans (Zhang et Rudnicky, 2001), en partant de la prémisse que les hypothèses de reconnaissance correctement reconnues sont plus grammaticales que celles incorrectement reconnues, on utilise un analyseur sémantique (*parser*) afin d'extraire la mesure de confiance. Cet analyseur produit un arbre contenant des groupes de mots résultants du parsing. Ceci est une source d'information pour le calcul d'une mesure de confiance au niveau mot en considérant deux paramètres : le degré du repli du groupe de mots (en partant d'un *quatre-grammes*) et le mode de parsing qui indique si un mot a été analysé comme faisant partie d'un groupe et donne également la position du mot dans le groupe (milieu ou extrémités) et le deuxième. Les auteurs obtiennent

1. Le modèle de repli (*backing-off*), introduit par (Katz, 1987), utilise un modèle de langage plus général. Pour un modèle *n-gramme*, si un certain *n-gramme* (h, w) n'a jamais été observé dans le corpus d'apprentissage, on descend d'un degré et le modèle de niveau inférieur *(n-1)-gramme* est utilisé.

ainsi des performances qui dépassent celles d'autres mesures de confiance utilisées : le score acoustique normalisé, la probabilité *a posteriori* dans un graphe de mots calculée en utilisant soit le score du modèle de langage soit le score du modèle acoustique, la stabilité acoustique (sur une liste de N meilleures hypothèses, etc.). Les performances de la mesure de confiance ainsi proposée restent toutefois inférieures aux performances obtenues avec la probabilité *a posteriori* de mots dans un graphe (calculée en utilisant le modèle de langage et le modèle acoustique).

Dans (Palmer et Ostendorf, 2001), des connaissances sémantiques, comme l'appartenance d'un mot à une classe sémantique, appelée aussi *entité nommée*, comme une "localisation, organisation ou personne", ont été utilisées afin d'estimer la mesure de confiance d'un mot reconnu.

Une mesure de confiance peut être calculée non seulement pour les mots d'une phrase mais aussi pour la phrase entière. Par exemple, dans (Pao et al., 1998), les connaissances sémantiques ont été utilisées pour le calcul des mesures de confiance en définissant des classes sémantiques. Chaque classe reçoit un certain poids et une distance sémantique est calculée sur les N meilleures hypothèses à l'aide de l'algorithme d'alignement de Levensthein. Par exemple, pour une classe qui regroupe des villes dans une application qui donne les prévisions météo, la substitution d'un mot de cette classe par un autre mot de la classe produit une distance sémantique très grande. Ainsi, les auteurs utilisent les poids des classes sémantiques de la meilleure solution ainsi que les distances sémantiques de trois premières solutions comme paramètres pour calculer une mesure de confiance à l'aide d'arbres de décision. Dans (San-Segundo et al., 2001), des paramètres prédictifs du modèle de langage, comme ceux cités au dessus, mais aussi des paramètres liés à l'analyseur sémantique en concepts utilisé par les auteurs, sont utilisés pour estimer une mesure de confiance au niveau de la phrase. Dans (Estève et al., 2003), l'estimation de la mesure de confiance au niveau de la phrase se base sur le fait que les événements non vus dans le corpus d'apprentissage lors de la construction du modèle de langage sont mal modélisés. Ainsi, pour une phrase donnée, la mesure de confiance est estimée en calculant le rapport entre le nombre de n-grammes présents dans la phrase et ayant été observés dans le corpus d'apprentissage du modèle de langage et le nombre total des n-grammes dans la phrase.

Discussions

On observe une polyvalence des paramètres linguistiques en ce qui concerne l'estimation de mesures de confiance tant au niveau mot qu'au niveau phrase. A la différence des paramètres acoustiques, certains paramètres linguistiques comme le repli ou les *entités nommées* peuvent être utilisés aussi bien au niveau mot qu'au niveau phrase. Par exemple, dans (Uhrik et Ward, 1997), le repli donne de très bonnes performances au niveau phrase mais des performances moindres au niveau mot. Ceci s'explique par une confusion pour savoir exactement quel mot est incorrect dans le cas des insertions et omissions des mots. Une adaptation est proposée afin de définir des groupes de mots qui peuvent contenir des erreurs potentielles ce qui augmente considérablement les

performances. Les *entités nommées* constituent elles aussi une technique qui peut être employé pour le calcul des mesures de confiance au niveau mot (Palmer et Ostendorf, 2001) et au niveau phrase (Pao et al., 1998).
Une observation intéressante est faite dans (San-Segundo et al., 2001), qui utilise le repli ainsi que le score de modèle de langage comme paramètres linguistiques et les compare à des paramètres acoustiques, comme la stabilité acoustique (pour une liste de N-meilleures hypothèses) ou le score acoustique normalisé par trame. Sur un corpus collecté à l'aide du système *CU Communicator* et utilisant un *multi-layer perceptron* pour combiner les différents paramètres, les auteurs observent de meilleures performances pour les paramètres linguistiques comparés aux paramètres acoustiques permettant une meilleure détection des mots incorrects. Il est à noter toutefois que les meilleurs résultats sont obtenus en combinant les deux types de paramètres. Des résultats similaires ont été observés dans (Zhang et Rudnicky, 2001) avec la précision que la probabilité *a posteriori*, qui peut être vue comme une combinaison des paramètres acoustiques (le score acoustique) et des paramètres de langage (le score de modèle de langage), dépasse les performances des paramètres linguistiques ou acoustiques. Cette observation va dans le sens de nombreuses publications qui s'accordent à dire que la probabilité *a posteriori* calculée dans un graphe de mots peut constituer une meilleure mesure de confiance que les paramètres prédictifs.

2.2.3 Autres paramètres

Il existe aussi d'autres paramètres prédictifs utilisés dans l'estimation de la mesure de confiance calculés à partir des différents éléments du système de reconnaissance et qui ne peuvent pas être classés dans les deux catégories présentées précédemment.
Des paramètres comme la durée du mot ou du phonème, l'état du HMM dans lequel on se trouve peuvent être utilisés dans l'estimation de la mesure de confiance (Vergyri, 2000).
L'utilisation de la liste des N-meilleures hypothèses peut être aussi une source importante d'information (Guo et al., 2004; Gillick et al., 1997; Hazen et al., 2002). On peut estimer une mesure de confiance à partir de paramètres comme les scores des premières hypothèses, la différence de score entre la meilleure hypothèse et les suivantes, le nombre de fois qu'un mot apparaît dans les différentes hypothèses, etc. Ainsi, plus un mot est présent dans les hypothèses plus il a des chances d'être correct.
Les graphes de mots peuvent être aussi utilisés dans l'estimation de la mesure de confiance. S'ils sont utilisés principalement pour le calcul de la probabilité *a posteriori*, voir la section 2.3, il existe aussi d'autres paramètres tel que la densité de l'hypothèse, c'est-à-dire le nombre de transitions en concurrence dans l'intervalle temporel du mot w (Kemp et Schaaf, 1997; Wessel et al., 1999). Comme pour le cas précédent, plus il existe d'hypothèses du même mot dans un intervalle de temps, plus le mot a des chances d'être correct.

2.2.4 Combinaison de plusieurs paramètres prédictifs

Comme nous l'avons précisé, le paramètre prédictif "*idéal*" doit fournir une information suffisante afin de pouvoir départager les mots reconnus qui sont corrects de ceux reconnus à tort. Pour ce faire le recouvrement entre les distributions des classes correct/incorrect doit être le plus petit possible. Or ce n'est pas le cas des paramètres présentés car, comme le montrent de nombreuses études (Kemp et Schaaf, 1997; Schaaf et Kemp, 1997), le recouvrement des deux distributions est assez important même pour les meilleurs paramètres.
Afin d'essayer de contourner ce problème, la combinaison de plusieurs de ces paramètres semble être un moyen efficace pour obtenir de meilleures performances. Un classifieur permet de regrouper les différents paramètres prédictifs afin d'obtenir une seule mesure de confiance comprise entre 0 et 1 pour un mot w.

La **régression logistique** permet de combiner plusieurs paramètres prédictifs afin d'obtenir une mesure de confiance (Charlet et al., 2001). La probabilité qu'une hypothèse soit correcte étant donné les valeurs respectives des paramètres prédictifs acoustiques et linguistique utilisés est donnée par :

$$P(COR|MC_{acc}, MC_{lang}) = \frac{1}{1 + e^{-(a_0 + a_1 \cdot MC_{acc} + a_2 \cdot MC_{lang})}} \tag{2.11}$$

Les paramètres a_0, a_1 et a_2 sont estimés de façon à minimiser l'entropie croisée (voir 2.1.3 sur un corpus de développement.
La régression logistique peut être utilisée aussi pour évaluer une seule mesure de confiance. Cette mesure de confiance doit permettre d'estimer la probabilité qu'une hypothèses soit correcte étant donné la valeur de la mesure de confiance. Cette probabilité, noté $P(COR|MC)$, peut être approximée à l'aide de la régression logistique en n'utilisant que deux paramètres de calibration :

$$P(COR|MC) = \frac{1}{1 + e^{-(a_0 + a_1 \cdot MC)}} \tag{2.12}$$

Les paramètres a_0 et a_1 sont estimés de façon à minimiser l'entropie croisée sur un corpus de développement.

Une autre méthode qui permet d'obtenir une mesure de confiance pour une hypothèse w est **l'interpolation linéaire**.

$$MC(w) = \sum_{n=1}^{N} c_n MC_n(w), \text{ avec } \sum_{n=1}^{N} c_n = 1 \tag{2.13}$$

Dans (Guo et al., 2004), les coefficients c_n ont été appris de manière à minimiser le taux d'erreur sur un corpus de type *Switchboard*. L'optimisation de ces coefficients a été faite en utilisant une procédure de validation croisée. Cette procédure consiste en un découpage du corpus en plusieurs parties, trois dans ce cas, et l'utilisation de deux corpus pour le calibrage des coefficients et du troisième pour le calcul du taux d'erreur. La

procedure tente ainsi de trouver le vecteur de coefficients qui obtient les meilleures performances sur les trois corpus de tests. Les auteurs combinent une mesure de confiance liée à la notion d'information mutuelle inter-mots (pour un mot elle se calcule comme la moyenne de l'information mutuelle du mot et des autres mots de la phrase) et une mesure de confiance utilisée plus fréquemment, la probabilité *a posteriori*, dont nous parlons au 2.3. Ces deux mesures de confiance étant indépendantes elles peuvent être combinées, l'interpolation linéaire permettant ainsi d'obtenir une amélioration des performances en termes de taux d'égale erreur (voir 2.1.1) de près de 10% par rapport à l'utilisation des deux mesures séparément.

Les **arbres de décision** (Kemp et Schaaf, 1997; Zhang et Rudnicky, 2001; Fu et Du, 2005; Kobus, 2006) sont souvent utilisés dans la reconnaissance de parole afin de prendre une décision binaire *correct/incorrect* sur les hypothèses de reconnaissance. Pour calculer la mesure de confiance d'un mot on y associe un vecteur de paramètres prédictifs. Les critères de décision associés à chaque nœud et les valeurs des paramètres déterminent, grâce à une décision binaire, le chemin à parcourir dans l'arbre de décision pour arriver à une feuille de l'arbre. Chaque feuille est associée à une probabilité que l'hypothèse soit correcte. Du fait de leur construction, les arbres de décision permettent d'avoir une vision sur les interactions des différents paramètres ainsi que sur la contribution de chacun d'entre eux.

Les **réseaux de neurones** (Kemp et Schaaf, 1997; San-Segundo et al., 2001; Charlet et al., 2001) prennent en entrée un vecteur constitué d'un certain nombre de paramètres prédictifs et permettent d'obtenir en sortie une mesure de confiance pour un mot w. En comparaison avec des méthodes de combinaison des paramètres comme l'interpolation linéaire ou les arbres de décision, les réseaux de neurones sont les classifieurs qui donnent les meilleures combinaisons des paramètres.

Une autre technique utilisée pour la combinaison des paramètres est la classification par **SVM** (*support vector machine*) (Zhang et Rudnicky, 2001). Il s'agit de délimiter au mieux deux nuages de points représentant les deux classes *correct/incorrect*. Pour ce faire les SVM utilisent des *fonctions kernel* qui doivent être évaluées afin d'obtenir une mesure de confiance (pour plus de détails voir (Burges, 1998), qui réalise une introduction très explicite de cette approche). Certaines fonctions apportent une amélioration des performances par rapport aux arbres de décision et aux réseaux de neurones. Si pour les réseaux de neurones le calibrage des paramètres utilisés est aisé, il n'en est pas de même pour les SVM. Les *fonctions kernel* ne sont pas très robustes et le moindre changement des paramètres peut produire un résultat sensiblement différent (Zhang et Rudnicky, 2001).

Il existe également d'autres classifieurs comme les modèles linéaires généralisés (*Generalized Linear Model*) présentés dans (Gillick et al., 1997; Siu et al., 1997; Siu et Gish, 1999) ou les méthodes de boosting (Moreno et al., 2001).

Comme nous l'avons vu, les SVM représentent un très bon classifieur avec des per-

formances supérieures aux arbres de décision et aux réseaux de neurones. Néanmoins, ils restent difficile à manier du fait que leur performances dépendent de la *fonction kernel* choisie mais aussi de l'optimisation des paramètres de cette fonction, opération qui peut s'avérer difficile. D'un autre coté, les réseaux de neurones et les arbres de décision utilisant des paramètres prédictifs de langage et acoustiques produisent de moins bonnes performances comparé à la probabilité *a posteriori* sur un graphe de mots (Zhang et Rudnicky, 2001).

2.3 Probabilité *a posteriori*

Comme nous l'avons montré au chapitre 1, un SRAP utilise le critère de *maximum a posteriori* pour trouver la séquence de mots $\hat{W}$ qui maximise la probabilité *a posteriori* $P(W|X)$ étant donné le signal acoustique X (voir l'équation 1.1). En théorie, la probabilité *a posteriori* est une très bonne mesure de confiance, mais cette formule est très difficile à utiliser du fait de la grande variabilité dans l'ensemble de départ des observations acoustiques. Pour cela, en pratique, un SRAP essaie de trouver la séquence de mots $\hat{W}$ qui maximise le produit $P(W) \cdot P(X|W)$ obtenu à l'aide de la formule de Bayes (voir l'équation 1.3). On observe que le terme $P(X)$ a été omis car il est indépendant de la séquence W. Cette méthode de calcul des scores acoustiques explique pourquoi ces scores sont inadaptés en tant que mesure de confiance.
Donc, pour calculer la probabilité *a posteriori* il suffit de normaliser par $P(X)$. En théorie, ce terme ce calcul de la manière suivante :

$$P(X) = \sum_{hyp} P(hyp) \cdot P(X|hyp) \tag{2.14}$$

où hyp représente une hypothèse possible du signal X, et la somme doit se faire sur toutes les hypothèses possibles, ce qui inclut toute combinaison possible de mots, phonèmes, bruits et autres événement acoustique. Il est évident que sans une contrainte particulière, $P(X)$ est très difficile à estimer de manière exacte. En pratique certaines contraintes doivent être imposées ainsi que des méthodes approximatives d'estimation du $P(X)$.
Dans une première catégorie on rencontre les modèles de type *filler* ("*remplissage*") présentés dans (Cox et Rose, 1996; Kampari et Hazen, 2000; Young, 1994). Ces approches peuvent obtenir des performances assez raisonnables. Dans une autre catégorie on rencontre les probabilités basées sur des listes de N-meilleures solutions et ensuite les probabilités basées sur les graphes de mots, qui incluent aussi les réseaux de confusion, calculés suite à des étapes de post traitement. La première catégorie, correspondant aux approches basées sur de modèles de type *filler*, ne fait pas l'objet de travaux dans cette thèse et ne sera détaillée. En revanche, dans cette partie, nous allons détailler les probabilités *a posteriori*, leur calcul et leur application, basées sur les graphes de mots, les listes de N-meilleures hypothèses et les réseaux de confusion.

2.3.1 Approximation par graphes de mots

Un système de reconnaissance automatique de la parole peut produire en sortie un graphe de mots. Comme le montrent les auteurs dans (Wessel et al., 1998, 1999, 2000, 2001; Kemp et Schaaf, 1997; Metze et al., 2000; Goel et al., 2001; Soong et al., 2004) il est possible d'approximer la probabilité *a posteriori* sur un espace plus restreint de solutions qui est le graphe de mots. Lors du décodage, la mémoire utilisée par le système de reconnaissance ainsi que la rapidité d'exécution de l'algorithme sont des facteurs importants à prendre en considération. Pour cela, comme il a été expliqué au chapitre précédent, des techniques d'élagage et de *beam search* sont employées lors de la génération du graphe de mots. De ce fait, le graphe de mots ne peut pas contenir toutes les hypothèses possibles. Néanmoins, l'approximation du terme $P(X)$ sur le graphe de mots reste correcte d'autant plus que le graphe de mots ne contient que les hypothèses les plus probables qui sont dominantes dans le calcul de $P(X)$.
Un graphe de mots G pour un vecteur d'observations acoustiques X est donc constitué de transitions auxquelles sont associées un mot w, son score acoustique $S(w)$ un instant de début et un instant de fin. Dans tout graphe de mots, il existe deux états particuliers : le premier est l'état de début du graphe qui correspond à l'instant de début de l'observation acoustique et l'état de fin qui correspond à l'instant de fin de l'observation acoustique. Tout chemin dans le graphe qui relie ces deux états s'appelle un chemin complet et représente une hypothèse de l'observation acoustique. Si on considère le chemin complet $C = T(w_1, d_1, f_1), T(w_2, d_2, f_2), \ldots, T(w_n, d_n, f_n)$ formé de n transitions, la probabilité de la séquence de mots représentée par ce chemin se calcule de la manière suivante :

$$P(C|G) = \prod_{i=1}^{n} S(w_i)_{d_i}^{f_i} \cdot P(w_i|h_i) \tag{2.15}$$

où h_i représente l'historique du mot w_i pour le chemin C, $P(w_i|h_i)$ représente la probabilité du modèle de langage calculée avec un modèle de type *n-gramme*et $S(w_i)_{d_i}^{f_i}$ représente le score acoustique associé à la transition T ayant pour instant de début d_i et de fin f_i et portant le mot w_i.
En utilisant les notations ci-dessus, pour une transition T du graphe de mots G, la probabilité *a posteriori* $P(T|G)$ se calcule comme étant le rapport entre la probabilité de tous le chemins dans G passant par T et la probabilité de tous le chemins dans G :

$$P(T|G) = \frac{\sum_{C \in G, T \subset C} P(C|G)}{\sum_{C \in G} P(C|G)} \tag{2.16}$$

où $C \in G$ représente un chemin complet dans le graphe G et $T \subset C$ montre que le chemin C contient la transition T. La probabilité *a posteriori* $P(T|G)$ peut être calculée de manière très efficace en utilisant l'algorithme *Forward-Backward* (Baum, 1972). Nous détaillons cet algorithme dans la partie 2.4.
La probabilité *a posteriori* d'une transition T peut être utilisée directement en tant que mesure de confiance pour le mot w porté par la transition. Il a été néanmoins montré (Wessel et al., 2001; Wessel et Ney, 2001) que cet usage est particulièrement inadapté et que les performances sont très limitées. La principale raison pour cela est le fait que dans un graphe de mots, à part la transition $T(w, d, f)$, il existe un nombre assez élevé

de transitions portant le mot w mais ayant des temps de début d et de fin f légèrement différents. La mesure de confiance du mot w est donc sous-estimée si on utilise seulement la probabilité *a posteriori* de la transition T. Il est alors important de prendre en compte les autres transitions portant le mot w et ayant un recouvrement temporel non-nul avec la transition T. Dans (Wessel et al., 2001) trois méthodes différentes sont proposées pour prendre en compte toutes les transitions $T(w, d \pm \varepsilon, f \pm \varepsilon)$ dans le calcul de la mesure de confiance pour le mot w.
Pour la première méthode le calcul de la mesure de confiance, appelée C_{sec}, pour la transition $T(w, d, f)$ s'effectue en sommant les probabilités *a posteriori* de toutes les transitions portant le même mot et ayant un recouvrement temporel non-nul avec la transition T (les transitions portant le même mot n'ont pas forcément un recouvrement temporel non-nul entre elles). Toutefois, avec cette méthode, la somme des probabilités *a posteriori* C_{sec} des différents mots pour une trame donnée entre la trame 0 et la trame de fin du graphe de mots n'est plus égale à 1. Malgré de meilleures performances pour C_{sec} par rapport à l'utilisation des probabilités *a posteriori* calculées avec l'algorithme *Forward-Backward*, le fait de ne pas normaliser cette mesure de confiance peut avoir des influences négatives sur la mesure de confiance. La normalisation devrait se faire de manière à ce que la somme des probabilités *a posteriori* C_{sec} des différents mots pour une trame donnée soit égale à 1. La deuxième méthode proposée essaie de contourner ce problème de normalisation. Le calcul de la probabilité *a posteriori* de la transition T est restreint à la sommation des probabilités *a posteriori* des transitions portant la même hypothèse de mot et dont les supports temporels, incluant celui de la transition T, doivent avoir une trame commune. De cette manière la somme de ces probabilités cumulées pour différentes hypothèses de mot à un instant de temps donné (compris entre 0 et l'instant de fin de l'énoncé) est égale à 1. Ainsi, la nouvelle mesure de confiance pour la transition T, appelée C_{med}, est calculée en sommant sur toutes les transitions portant le mot w qui ont un chevauchement temporel avec la trame médiane de la transition T. Les résultats montrent des performances comparables à C_{sec} et l'absence de normalisation dans le calcul de celle-ci semble ne pas avoir d'influence sur les performances. La troisième méthode se propose de déterminer si le choix de la trame d'intersection à une influence sur la mesure de confiance. Le calcul pour la mesure de confiance, appelée C_{max}, se fait en sommant les probabilités *a posteriori* non seulement pour la trame médiane, comme expliqué pour C_{med}, mais pour toutes les trames du support temporel de la transition T et en choisissant la valeur maximale de ces sommes comme étant la valeur de C_{max}. Cette nouvelle mesure de confiance engendre des performances légèrement meilleures que les deux précédentes. L'évaluation des performances des trois mesures de confiance décrites a été effectuée sur plusieurs corpus de test (ARISE (dialogues homme-machine via un service téléphonique d'information sur les horaires de train), Verbmobil (parole spontanée), Broadcast News (journaux télévisé et radio), NAB (articles lus à partir de différents journaux)) en évaluant le taux d'erreur de confiance (CER). Le taux de référence a été défini comme le rapport entre la somme de nombre d'insertions et de substitutions divisée par le nombre total de mots reconnus.

La probabilité *a posteriori* peut être utilisée aussi dans le domaine de l'apprentissage non-supervisé des modèles acoustiques. Dans (Wessel et Ney, 2001), l'utilisation des

probabilités *a posteriori* en tant que mesure de confiance permet de réduire considérablement la taille du corpus d'apprentissage (les transcriptions manuelles) avec une dégradation moindre des performances du système en termes de taux d'erreur mot.

2.3.2 Approximation par liste de N meilleures hypothèses

Comme pour les graphes de mots, la probabilité *a posteriori* d'un mot w peut également être calculée sur une liste de N meilleures hypothèses (liste N *best*) (Wessel et al., 1999, 2001; Rueber, 1997). Chaque énoncé d'une liste de N *best* est seulement une séquence de mots sans aucune information temporelle sur les temps de début et de fin de chaque mot. Comme décrit dans la section précédente (voir 2.3.1), la relaxation des frontières des hypothèses de mots dans un graphe est très importante pour le calcul d'une mesure de confiance fiable. Un des principaux avantages d'une liste N *best* est alors l'absence de l'information temporelle sur les hypothèses de mots de chaque énoncé. La mesure de confiance d'une hypothèse de mot peut être calculée en se basant seulement sur la position des mots dans les énoncés. Néanmoins, la notion de la position du mot dans la phrase n'est pas très bien définie. Tout d'abord les phrases de la liste N *best* n'ont pas forcement la même longueur. Même dans le cas contraire, les mots ne peuvent pas être comparés directement à cause des erreurs d'insertion et d'omission. Pour ces raisons, les phrases de la liste doivent être alignées en utilisant l'algorithme de Levensthein (Levensthein, 1966). L'algorithme vise à minimiser la somme des insertions, omissions et substitutions dans la comparaison entre deux phrases.
Pour chaque mot w_m se trouvant à la position m dans la phrase w_1^M, l'algorithme peut lui faire correspondre un mot v dans n'importe quelle autre phrase $v_1^{M_1}, \ldots, v_1^{M_N}$ de la liste. On écrit cela de la manière suivante : $v = L_m(w_1^M, v_1^{M_n})$. La fonction $L_m(w_1^M, v_1^{M_n})$ donne le mot v dans la phrase $v_1^{M_n}$ qui correspond au mot w_m de la phrase w_1^M selon l'alignement de Levensthein. Utilisant ces notations, la probabilité *a posteriori* du mot w_m dans la phrase de référence w_1^M se calcule de la manière suivante :

$$p(w_m|X, w_1^M, L) = \frac{\sum_{n=1}^{N} p(X|v_1^{M_n})^{\alpha} p(v_1^{M_n})^{\beta} \delta(w_m, L_m(w_1^M, v_1^{M_n}))}{\sum_{n=1}^{N} p(X|v_1^{M_n})^{\alpha} p(v_1^{M_n})^{\beta}} \tag{2.17}$$

où la fonction de Kronecker δ est égale à 1 si les deux arguments sont identiques et à 0 autrement. α et β sont des facteurs d'échelle qui doivent être optimisés sur un corpus de développement.
Dans (Wessel et al., 2001), les auteurs ont montré que les performances des probabilités *a posteriori* calculées sur une liste N *best* peuvent être similaires ou moins bonnes que celles des probabilités *a posteriori* calculées sur des graphes de mots en fonction de la tâche choisie. Les performances sont mesurées en comparant le taux d'erreur de confiance. Sur les corpus ARISE et NAB, les performances des deux types de probabilités *a posteriori* sont équivalentes alors que sur Verbmobil et Broadcast News les probabilités *a posteriori* sur les graphes de mots, et plus particulièrement C_{max} (voir section précédente), ont des performances très supérieures aux probabilités *a posteriori* calculées sur la liste N *best*. Plusieurs explications ont été avancées à ce sujet. Ainsi, sur ARISE, l'équivalence des performance peut être attribuée à une longueur moyenne de l'énoncé

très courte, d'environ 3.4 mots par énoncé. Même si les corpus NAB et Broadcast News ont des caractéristique similaires (taille du lexique équivalente, longueur moyenne des énoncé très grande) les performances sont loin d'être similaires. Une explication peut résider dans le fait que le premier corpus est enregistré dans des conditions audio nettement supérieures et les énoncés sont lus. Ainsi, du fait d'une meilleure qualité des modèles acoustique et de langage la distribution de probabilité est plus discriminante entre les différentes hypothèses de mot. Ceci mène à l'utilisation de moins d'énoncés dans la liste de *N best* pour le calcul de la mesure de confiance. C'est aussi la combinaison entre la taille de la liste *N best* et les modèles acoustique et de langage qui peut expliquer les moins bonnes performances pour les corpus Verbmobil et Broadcast News.
Dans (Wessel et al., 1999, 2001) il a aussi été montré que la probabilité *a posteriori*, qu'elle soit basée sur les graphes de mots ou sur une liste de *N best*, engendre des performances nettement supérieures, en termes de taux d'erreur de confiance, par rapports aux mesures de confiance, comme la stabilité acoustique ou la densité de l'hypothèse, décrites au 2.2.
Pour une liste *N best*, on serait tenté de dire que plus elle est grande meilleures sont les performances. Ceci n'est pas forcement vrai, ce qui est démontré aussi par une observation intéressante faite par les auteurs dans (Wessel et al., 2001) qui montre la dégradation des performances sur la liste *N best* pour une $N = 1000$ (les autres valeurs de N sont 100, 200 et 300). De plus, l'utilisation des listes *N best* très grandes n'est pas très efficace en termes de complexité et temps de calcul et des ressources nécessaires. Une analyse détaillé des listes *N best* (Wessel et al., 2001) montre aussi que des mots dont les supports temporels sont éloignés sont parfois mis en correspondance. L'algorithme de Levensthein mène parfois à des alignements de mots "*peu raisonnables*" ce qui peut conduire à des problèmes supplémentaires dans le calcul de la probabilité *a posteriori*. L'information sur le temps de début et de fin des mots contenue dans les graphes de mots peut donc s'avérer importante et nécessaire pour une meilleure estimation de la probabilité *a posteriori*.

2.3.3 Approximation par réseaux de confusion

Comme nous l'avons expliqué au 2.3.1, l'utilisation directe de la probabilité *a posteriori* des transitions portant des mots dans un graphe de mots en tant que mesure de confiance du mot peut poser des problèmes de fiabilité. Comme montré ceci est du à l'alignement temporel des différentes hypothèses du même mot et pour cela différentes méthodes qui essaient de calculer une meilleure mesure de confiance sur les mots du graphe ont été présentées dans la littérature et certaines détaillées au 2.3.1. Les réseaux de confusion, de part leur construction, évitent ce problème d'alignement car dans une classe de mots, qui est associée en fait à un intervalle temporel, il ne peut y avoir qu'un seul arc portant un mot w et donc une seule probabilité *a posteriori* du mot dans cet interval du temps. Ce problème est néanmoins présent lors de la construction des CNs, mais cette problématique sera détaillée au chapitre 5. Dans cette partie nous allons seulement décrire l'utilisation des réseaux de confusion, et plus précisément des probabilités *a posteriori* des mots, dans le calcul de mesures de confiance en présentant

quelques exemples décrits dans la littérature.
Dans (Evermann et Woodland, 2000a) par exemple, les auteurs réalisent une comparaison entre les performances des probabilités *a posteriori* dans les graphes de mots et dans les réseaux de confusion. Pour les graphes de mots, une technique de *rescoring* est utilisée. Dans une première étape les probabilités *a posteriori* de transitions du graphe sont calculées à l'aide de l'algorithme *Forward-Backward* et ensuite ces probabilités sont combinées avec les scores acoustiques des transitions. Un simple algorithme A^* est utilisé pour calculer la meilleure solution. Les résultats de cette solution en termes de WER et de NCE(entropie croisée normalisée, voir 2.1.3) sont comparés aux performances de la *consensus hypothesis* sur un corpus de Broadcast News et un corpus contenant des conversations téléphoniques. Les réseaux de confusion se révèlent être plus robustes que la technique de *rescoring* utilisée sur les graphes de mots en produisant des améliorations similaires en termes de WER et SER sur les deux corpus. Si en termes de WER les tests sont assez concluants, en termes de NCE, les résultats montrent que les deux approches (les probabilités *a posteriori* C_{sec} (Wessel et al., 2001) basées sur les graphes de mots et les probabilités *a posteriori* des mots dans les réseaux de confusion) donnent des résultats comparables. Les modèles acoustiques utilisés (quinphone ou triphone) et la taille des graphes de mots font pencher la balance dans un sens ou dans un autre. L'utilisation d'un arbre de décision pour le calcul des scores de confiance améliore les résultats pour les deux approches.
Dans (Xue et Zhao, 2006) on propose de comparer les performances en termes de taux d'erreur de confiance sur les annotations (voir 2.1.2) de trois classifieurs : SVM, arbre de décision et *random forests* (forêts aléatoires). La nouveauté de ces travaux réside dans l'utilisation de nouveaux paramètres issus des CNs dans les *random forests*. Les *random forests* (Breiman, 2001) sont constitués d'un large nombre d'arbres de décision. Afin de classifier un objet, celui-ci est traité par chaque arbre de la forêt, la classification avec le plus grand nombre de votes étant choisie. Une caractéristique importante de ce classificateur est celle de pouvoir estimer quels paramètres sont les plus importants dans le processus de classification.
Dans ces travaux, les *random forests* utilisent des paramètres issus du décodeur de parole (les scores acoustiques et de langage, le score *a posteriori* local d'un mot) et des paramètres issus des CNs (la probabilité *a posteriori* des mots, la probabilité *a posteriori* conditionnelle étant donné le contexte, l'entropie pour les CNs). La probabilité *a posteriori* conditionnelle est calculée étant donné un contexte de type bigramme ($P(w_i|w_{i-1}, X)$) ou de type trigramme ($P(w_i|w_{i-1}, w_{i-2}, X)$). L'entropie pour les CNs mesure les différences entre les probabilités *a posteriori* des mots de la même classe et elle est calculée comme la somme sur tous les mots d'une classe du produit entre la probabilité *a posteriori* du mot et son logarithme. Les évaluations sont réalisées sur des énoncés collectés à partir d'une application de télé-médecine. Les résultats montrent, d'un coté, de meilleures performances des *random forests* par rapport aux SVM et aux arbres de décision, et d'un autre coté l'importance de l'entropie et des probabilités *a posteriori* dans les *random forests*. Le fait d'enlever seulement un des deux paramètres engendre des baisses de performances en terme d'erreur d'annotation. Du fait d'un certain degré de corrélation entre les deux paramètres, l'absence de l'entropie et de la probabilité *a posteriori* en même temps engendre la plus grande baisse des performances par rapport à

l'absence d'autres paramètres.
La combinaison des systèmes est une technique qui est devenue très populaire ces dernières années car elle permet l'obtention d'une solution ayant de meilleures performances que n'importe quelle solution des systèmes de départ. Dans (Evermann et Woodland, 2000b), les auteurs proposent une amélioration de la technique ROVER. A la base, celle-ci utilise des séquences de mots de plusieurs systèmes et une technique de programmation dynamique pour réaliser un alignement des séquences. Grace à cet alignement une décision est prise entre les mots alignés soit par un simple vote, soit en tenant compte aussi des scores de confiance. La limite de la technique ROVER réside dans l'utilisation des séquences de mots et dans la procedure de décision. Ainsi, seule une hypothèse de mot choisie par un des systèmes peut faire partie de la solution finale. L'amélioration proposée par les auteurs permet d'utiliser l'alignement des CNs produits par les systèmes afin de décider sur les hypothèses de mots. Cette technique permet d'obtenir des améliorations en termes de WER ainsi que d'entropie croisée normalisée (*Normal Crossed Entropy - NCE*). Les évaluations ont été effectuées sur des corpus de type Switchboard et CallHome.
Comme nous l'avons décrit, les probabilités *a posteriori* des mots dans un réseau de confusion peuvent être utilisées de manière efficace en tant que mesure de confiance. Une légère tendance des probabilités *a posteriori* des mots de surestimer la probabilité que le mot soit correct à été observée dans (Evermann et Woodland, 2000b). Cet effet a été plus prononcé sur des systèmes ayant des taux d'erreur mot plus grands et sur des graphes de mots de petite taille. Des techniques pour compenser cette surestimation sont proposées comme les arbres de décision ou les réseaux de neurones qui calculent les scores de confiance.
Dans les travaux présentés nous utilisons également les probabilités *a posteriori* de mots dans un réseau de confusion pour le calcul des scores de confiance. Le calcul de la probabilité que le mot soit correct étant donné sa probabilité *a posteriori* sera effectué en utilisant la régression logistique. Les coefficients nécessaires au calcul du score de confiance (voir la section 2.2.4) sont appris sur un corpus de développement. Les scores de confiance ainsi calculés sont ensuite utilisés dans le filtrage de la meilleure solution des CNs en appliquant un seuil sur la valeur des scores de confiance. La séquence de mots ainsi obtenue est ensuite utilisée dans la construction des différentes stratégies d'interprétation qui améliorent les performances du système de dialogue utilisé (voir section 3.1). Les évaluations de l'utilisation de ces scores de confiance sont présentées au dernier chapitre.

2.4 Calcul de la probabilité *a posteriori* sur les graphes de mots

Le calcul de la probabilité *a posteriori* d'une transition dans un graphe de mots est réalisé à l'aide de l'algorithme *Forward-Backward* (Baum, 1972). Cet algorithme a été décrit de manière détaillé dans (Rabiner, 1989; Rabiner et Juang, 1993; Jelinek, 1997) afin de répondre à certaines questions soulevées par l'utilisation des HMM. Toutefois, dans notre cas, l'émission des observations est attachée au transitions et non aux nœuds comme c'est le cas pour les HMM dans (Rabiner, 1989). Nous avons donc du adapter

les formules de calcul décrites dans (Rabiner, 1989) pour le calcul de la probabilité *a posteriori* sur un graphe de mots. Le détail de ces formules est présenté dans la section A.1 de l'annexe A.
Le calcul de la probabilité *a posteriori* de la transition t, se fait en deux temps à travers une procédure "*forward*" (qui définit la variable $\alpha(t)$) qui réalise une passe avant dans le graphe sur les chemins partiels jusqu'à la transition t et une procédure *backward* (qui définit la variable $\beta(t)$) qui réalise une passe arrière sur les chemins partiels partant de la transition t jusqu'à l'état de fin de graphe. Toutefois, le calcul des variables définies par ces procédures pose des problèmes numériques de type *underflow* car on multiplie des variables inférieures à 1. Ces problèmes sont liés à la capacité limitée des ordinateurs à représenter un nombre positif, très proche de 0, avec beaucoup de décimales après la virgule.
Une solution couramment employée dans la littérature est de ne pas prendre en compte les transitions pour lesquelles le calcul de la probabilité *a posteriori* est impossible. Un élagage de ces transitions est alors obligatoire. De plus, cet élagage du à un problème d'*underflow* sur une des variables $\alpha(t)$ ou $\beta(t)$ (les variables sont considérées comme égales à 0, et donc la probabilité *a posteriori* de la transition est égale à 0) rend le calcul des probabilités *a posteriori* approximatif. Par exemple, la variable $\alpha(t) = 0$ implique un élagage de la transition t. Le calcul de $\alpha(t')$, avec t' un successeur de t dans le graphe (l'instant de fin de t est égal à l'instant de début de t'), dépend de $\alpha(t)$ (voir la section A.1). L'élagage de la transition t implique alors que l'ensemble des chemins partiels arrivant dans t est éliminé du calcul de la probabilité *a posteriori* sur t'. Cette méthode est approximative et force un élagage des transitions pour des raisons de calcul numérique. Et cela sans qu'on puisse mesurer son influence sur les performances de graphes de mots.
Une deuxième solution est proposée dans (Rabiner, 1989) et consiste en l'introduction des coefficients de normalisation des variables α et β. Nous avons choisi cette deuxième méthode car l'introduction de ces coefficients permet un calcul exact de la probabilité *a posteriori* en éliminant les problèmes d'*underflow* et sans qu'aucun élagage soit nécessaire. Néanmoins, dans (Rabiner, 1989) cette méthode est décrite dans le contexte de l'utilisation de l'algorithme *Forward-Backward* sur les HMM. Par conséquence, les coefficients sont définis pour une normalisation au niveau de la trame. Leur définition ne peut pas s'appliquer directement aux graphes de mots car les variables α et β sont calculées au niveau de la transition qui s'étend sur plusieurs trames. Nous proposons donc une nouvelle normalisation des variables qui couvre l'ensemble des trames du support temporel de la transition. Nous avons adapté les formules de calcul de α et β pour intégrer cette nouvelle normalisation. Ces formules sont détaillées dans la section A.2 de l'annexe A.

2.5 Conclusions

Dans ce chapitre nous avons tout d'abord présenté les différentes métriques utilisées pour l'évaluation des mesures de confiance. Nous avons ensuite présenté deux catégories de paramètres utilisés dans le calcul des mesures de confiance : les paramètres

prédictifs et la probabilité *a posteriori*. Dans le cas des paramètres prédictifs nous avons détaillé les différents types de paramètres (acoustiques, linguistiques) ainsi que les techniques les plus courantes de combinaison de paramètres. Pour la probabilité *a posteriori* nous avons présenté trois méthodes d'estimation à partir des graphes de mots, liste de *N best* et réseaux de confusion.
Les travaux de cette thèse n'essaient pas de faire une comparaison des différentes mesures de confiance ou d'étudier les performances des différents classifieurs sur les réseaux de confusion. Comme nous l'avons évoqué à plusieurs reprises, la probabilité *a posteriori* peut constituer une très bonne mesure de confiance. De plus, il a été montré aussi que la probabilité *a posteriori* des mots dans un réseau de confusion constitue une meilleure mesure de confiance que la probabilité *a posteriori* des mots dans un graphe (Evermann et Woodland, 2000a). Nous avons donc décidé d'utiliser la probabilité *a posteriori* des mots dans les réseaux de confusion en tant que mesure de confiance. Afin de faciliter les calculs, nous avons choisi la régression logistique pour calculer la probabilité conditionnelle qu'un mot soit correct étant donné sa probabilité *a posteriori*. La régression logistique s'est avérée être assez performante et la phase de calibrage de paramètres facile à mettre en œuvre.
Nous avons également détaillé l'algorithme *Forward-Backward*. Nous avons adapté les formules pour être utilisées dans le calcul de la probabilité *a posteriori* d'une transition dans un graphe de mot. L'implémentation de l'algorithme *Forward-Backward* pose des problèmes d'*underflow*. Une méthode de normalisation par trame des variables de l'algorithme a été proposée dans (Rabiner, 1989). Nous avons modifié cette méthode afin de définir une nouvelle normalisation par transition. Les formules qui intègrent cette nouvelle normalisation ont aussi étaient décrites. Nous avons choisi d'utiliser cette méthode de normalisation car elle permet un calcul exact de la probabilité *a posteriori* et ne nécessite aucun élagage des transitions.