Robustesse de la RAP à la parole expressive âgée vs. typique : contexte de commandes dans un habitat intelligent

Frédéric Aman[1, 2] Véronique Aubergé[1]
Michel Vacher[1]
(1) CNRS, LIG, F-38000 Grenoble, France
(2) Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
Frederic.Aman@imag.fr, Veronique.Auberge@imag.fr, Michel.Vacher@imag.fr

MOTS-CLÉS : Parole avec affect, appel de détresse, habitat intelligent pour la santé.

KEYWORDS: Expressive speech, distress call, Ambient Assisted Living.

1 Introduction

Dans le cadre de l'assistance aux personnes âgées vivant seules à domicile, notre objectif est de développer un système capable de détecter des énoncés de personnes âgées en détresse. L'étude que nous proposons ici vise à mettre en évidence le manque de robustesse des systèmes de reconnaissance automatique de la parole (RAP) avec de la voix prononcée de façon émue. Nous avons enregistré au laboratoire un corpus de phrases de détresse qui est un ensemble d'énoncés prononcés de façon neutre et émue actée dans un protocole d'élicitation. L'utilisation d'un modèle acoustique de la RAP adapté a permis de réduire en partie la baisse de performance rencontrée avec la parole émue.

2 Corpus de voix de détresse et caractérisation prosodique

Le corpus *Voix Détresse* a été constitué afin d'étudier l'impact d'une voix expressive sur un système de RAP en comparaison avec de la voix sans émotions (lecture). Il a été demandé aux locuteurs de lire 20 phrases de détresse de façon neutre. Puis, nous avons enregistré des émotions élicitées : une photographie représentant une situation où un personnage est en détresse a été associée à chaque phrase, et il a été demandé aux locuteurs de se mettre dans la peau des personnages et d'énoncer les phrases de façon très expressive. 20 locuteurs jeunes et 5 locuteurs âgés ont été enregistrés.

Nous avons réalisé une comparaison des paramètres prosodiques entre phrases neutres et émues. Nous avons observé pour la voix émue, par rapport à la voix neutre, une diminution du débit, une augmentation de la fréquence fondamentale, une diminution du *jitter*, une diminution du *shimmer*, et une augmentation du rapport harmonique sur bruit. Nous avons pu également observer que la nature même de la détresse pouvait prendre différentes formes selon la situation suggérée par la situation visuelle présentée pour l'élicitation, que ces formes étaient homogènes entre sujet, et que chaque forme était caractérisée par des valeurs spécifiques de paramètres prosodiques.

3 Expérimentation

Avec le système de RAP *Sphinx3*, nous avons utilisé un modèle acoustique de type HMM appris sur le corpus *BREF120*. Un modèle de langage spécifique aux phrases à reconnaître a été entraîné.

Les résultats du décodage montrent une dégradation importante du WER (taux d'erreurs de mots) entre les voix neutres et les voix émues pour les locuteurs jeunes (WER_{neutre}=9,27%, $WER_{émue}$=39,22%), ainsi que pour les locuteurs âgées (WER_{neutre}=18,82%, $WER_{émue}$=38,42%). Nous observons également une variabilité importante du WER entre locuteurs, notamment pour la voix émue (écart-type=16,95%), s'expliquant en partie par le fait que le corpus de parole émue utilisé était un corpus de parole actée, enregistré par des non professionnels. Globalement, les personnes ressenties comme étant les plus à l'aise pour jouer les situations de détresse sont les personnes ayant un WER le plus élevé pour les phrases émues. Nous pouvons donc craindre qu'en situation réelle, une situation de détresse exprimée de façon extrêmement émue sera très mal reconnue par un tel système de RAP, alors que c'est justement dans ce type de situation qu'il est primordial de pouvoir agir.

Nous avons réalisé une adaptation MLLR au locuteur sur le modèle acoustique générique *BREF120* selon 3 modalités : (1) adaptation à la voix neutre ; (2) adaptation à la voix émue ; (3) adaptation sans distinction neutre ou émue. Le décodage a été réalisé avec les différents modèles acoustiques adaptés aux locuteurs, et nous avons effectué une comparaison avec le modèle *BREF120* (voir table 1). Une ANOVA suivie d'un test de Tukey HSD a été réalisée sur les groupes *voix neutres* et *voix émues*. Les ANOVA montrent qu'il n'y a pas de différence significative entre les échantillons du groupe *voix neutres*, et qu'il existe une différence significative entre certains échantillons du groupe *voix émues*. Pour les voix émues, le test de Tukey HSD montre qu'il existe une différence de WER significative entre le modèle générique *BREF120* et les modèles adaptés à la voix émue *BREF120_MLLR_LOC_E* et *BREF120_MLLR_LOC_N+E*. En revanche, il n'y a pas de différence significative entre les autres modèles acoustiques. Ainsi, lors du décodage des phrases émues, nous voyons que l'adaptation au locuteur à partir de phrases neutres n'est pas suffisamment efficace pour améliorer significativement le WER. Il est donc nécessaire d'utiliser des modèles adaptés à la voix émue.

4 Conclusion

Nous avons observé une dégradation des performances jusqu'à 30% en utilisant un modèle générique appris sur le corpus BREF120. En adaptant ce modèle au locuteur avec les phrases émues l'amélio-

Groupe	BREF120	BREF120_MLLR_LOC_N	BREF120_MLLR_LOC_E	BREF120_MLLR_LOC_N+E
JN	9,27	7,80	11,03	7,21
JE	39,22	28,86	22,45	20,23
AN	18,82	17,06	16,48	15,88
AE	38,42	35,48	31,64	30,50
Moy. N	11,18	9,65	12,12	8,94
Moy. E	39,06	30,18	24,28	22,28

TABLE 1: *WER (%) en fonction des différents modèles acoustiques pour la voix des locuteurs jeunes (J) et âgés (A), avec une intonation neutre (N) ou émue (E).*

ration du WER est significative (15%), alors qu'elle ne l'est pas en utilisant les phrases neutres, ce qui montre l'importance d'adapter les modèles acoustiques à partir de données prononcées de façon émue pour une détection de phrases de détresse. Après adaptation, le WER des voix émues reste plus important que le WER des voix neutres avec le modèle générique BREF120, la différence étant de 13%. En outre, le corpus *Voix Détresse* enregistré contient de la voix émue actée, la parole émue spontanée étant beaucoup plus difficile à enregistrer. Dans un cas réel, il faudra donc s'attendre à des résultats encore plus dégradés.