Combinaison des sorties de plusieurs microphones pour l'amélioration de la reconnaissance vocale embarquée

Yannick Hallé, Lionel Delphin-Poulat

France Telecom R&D / TECH / SSTP / RVA
22307 Lannion Cedex France
yannick.halle, lionel.delphinpoulat@orange-ftgroup.com

ABSTRACT

In this paper, we propose to improve a speech recognition system intended to control in-car accessories such as navigation system. We focus our study on the speech signal recorded in the noisy in-car environment. The method proposed is to perform an acoustic analysis with embedded noise reduction and waveform processing on this signal, to train speech models by using bases of in-car audio recordings and to use the differences between the outputs from several microphones with the ROVER approach.
Keywords: Speech Recognition In-Car, Model Adaptation, ROVER

1. Introduction

Les interfaces à base de reconnaissance automatisée de la parole apparaissent comme étant un moyen intéressant pour améliorer les interactions entre le conducteur et les différents accessoires présents dans un véhicule, comme le navigateur ou un téléphone. La parole est en effet une manière naturelle d'interagir avec ces accessoires, tout en permettant aux yeux et aux mains de se concentrer sur la conduite en elle-même. L'intégration de ces interfaces à l'intérieur des véhicules intéresse donc les entreprises de ce secteur [1][2][3]. L'amélioration de ces interfaces de reconnaissance vocale embarquées se heurte à de nombreux problèmes techniques indépendamment de ceux propres à la reconnaissance vocale.

Parmi ces problèmes, la chute de performance causée par le bruit à l'intérieur d'un véhicule pendant la conduite est particulièrement importante. Ce bruit dépend d'un grand nombre de paramètres comme le type de véhicule, la vitesse, le fonctionnement d'accessoires ou l'ouverture des fenêtres. Lorsque la prise de son se fait avec un microphone distant, l'impact du bruit est d'autant plus important . Les travaux visant à réduire la chute de performances dans un véhicule peuvent être classés en quatre catégories [4], en considérant leur objectif initial : l'identification de paramètres du signal et de mesures de similarité résistante au bruit, l'amélioration du signal de parole bruitée, l'adaptation des modèles de reconnaissance à des environnements bruités et l'utilisation d'antennes acoustiques.

Dans cet article, nous considérons l'utilisation, par le conducteur, d'un système de reconnaissance vocale embarqué. Nous ajoutons à ce cadre de travail deux contraintes majeures. La première contrainte est l'utilisation d'un ou de plusieurs microphones distants pour la prise de sons. La deuxième contrainte est que le système de reconnaissance reçoit le signal de parole et lui seul (pas d'informations sur la vitesse du véhicule, le fonctionnement d'accessoires, etc.). Ce cadre ne nous permet pas de bénéficier des avancées obtenues par l'utilisation d'antennes acoustiques. De plus, cet article se focalise sur le signal enregistré par le microphone et ne traite pas le problème d'une transmission vers un serveur ou de l'ergonomie de l'application.

Nous avons décidé, premièrement, d'utiliser une analyse acoustique améliorée, l'analyse Aurora [5], qui comporte une technique de réduction de bruit et d'amélioration du signal et qui fournit de meilleures performances qu'une analyse classique par coefficients cepstraux MFCC (Mel Frequency Cepstral Coefficients) dans un environnement bruité, deuxièmement, d'adapter les modèles de reconnaissance à l'environnement véhicule en utilisant des enregistrements effectués dans des conditions réelles de conduite, et troisièmement, d'utiliser conjointement les signaux provenant de plusieurs microphones afin d'améliorer encore les performances.

Cet article se propose d'évaluer les améliorations apportées par ces trois points sur des énoncés enregistrés dans des conditions réelles de conduite.

2. Base de données et cadre expérimental

La base française Vodis [6] a été utilisée pour mener les différentes expérimentations. C'est une vaste base d'énoncés enregistrés, dans des conditions réelles de conduite, par deux cents locuteurs français à bord de trois voitures (VW Passat, Peugeot 406, Renault Laguna). Chaque locuteur a été enregistré simultanément avec un microphone de proximité (close-talk) et trois microphones AKG distants (far-talk) sur le tableau de bord, respectivement à gauche (LE), au milieu (MI) et à droite (RI) du conducteur. Les enregistrements ont été effectués en 16 bits linéaire à 11025Hz et les conditions d'enregistrement de chaque énoncé (vitesse du véhicule, fonctionnement de la radio et de la ventilation, ouverture des fenêtres,..) ont été renseignées dans les en-têtes de fichiers. La base Vodis regroupe plusieurs corpus d'énoncés auxquels ont participé tous les locuteurs : lettres, nombres,

suites de nombres, mots de commande, phrases spontanées, phrases de commande, numéros de téléphone, etc..

Les corpus utilisés pour mener nos tests de reconnaissance vocale sont le corpus des numéros de téléphone (avec 5 nombres de 2 chiffres par numéro), appelé "corpus T" qui combine une certaine variété phonétique et un vocabulaire fermé, et le corpus des mots de commande, appelé "corpus W", qui regroupe 80 mots ou groupes de mots de commandes simples ("décrochez", "c'est bon", "radio", "page précédente",...). Le corpus T contient 1136 enregistrements, avec un numéro de téléphone par enregistrement, et le corpus W contient 3547 enregistrements, avec un mot ou groupe de mots par enregistrement, le tout enregistré par les trois microphones far-talk (LE, MI et RI).

Deux corpus de phrases, les corpus R et S (respectivement "phrases de commande" et "phrases spontanées"), ont servi à effectuer l'adaptation des modèles. Cela représente 2608 énoncés, enregistrés par les trois micros far-talk. Afin d'éviter les biais liés à la présence d'un même locuteur à la fois dans les sessions d'énoncés servant à l'adaptation et les sessions d'énoncés testées, les locuteurs ont étés répartis en trois groupes (en essayant de respecter une répartition la plus homogène possible des conditions d'enregistrement entre ces trois groupes). Ainsi, les énoncés des corpus de tests appartenant à un des groupes de locuteurs seront testés avec un modèle adapté avec les énoncés des corpus d'adaptation des deux autres groupes de locuteurs. L'adaptation des modèles s'est fait par régression linéaire MLLR (Maximun Likelihood Linear Regression)[7].

Les expériences ont été menées en utilisant un moteur de reconnaissance vocale par modèles de Markov cachés, avec des modèles phonétiques dépendant du contexte. Les densités sont des mélanges de gaussiennes, avec au plus huit gaussiennes par densité.

L'analyse acoustique utilisée est l'analyse "Aurora Advanced Front-End for distributed speech recognition". Elle comporte une technique avancée de réduction de bruit et d'amélioration du signal. L'énergie, les 12 premiers coefficients et leurs dérivées temporelles premières et secondes sont utilisés. L'analyse est appliquée avec un décalage de trames de 10ms (les coefficients sont ensuite sous-échantillonnés à 16 ms) avec un signal sous-échantillonné à 8kHz (à partir du signal enregistré à 11025Hz). Associée à des modèles adaptés à l'environnement véhicule, elle améliore significativement la reconnaissance vocale par rapport à une analyse par MFCC [8].

3. Adaptation des modèles

L'adaptation des modèles au bruit à l'intérieur de l'habitacle s'effectue par régression linéaire grâce à des enregistrements des corpus R et S de la base Vodis, en supprimant le biais lié à la présence d'un même locuteur dans les bases de tests et d'adaptation comme expliqué dans le chapitre 2. Les taux d'erreur-mot présentés dans la suite de l'article proviennent de la concaténation des résultats de reconnaissance sur l'ensemble du corpus de test.

Micro. de test	Corpus T	Corpus W
Modèles phonétiques standard		
Micro LE	19.8% [18.8% - 20.9%]	4.6% [4.0% - 5.4%]
Micro MI	21.1% [20.0% - 22.1%]	5.4% [4.7% - 6.2%]
Micro RI	23.8% [22.7% - 24.9%]	7.0% [6.2% - 7.9%]
Modèles adaptés sur R et S LE-MI-RI		
Micro LE	11.4% [10.6% - 12.2%]	2.3% [1.9% - 2.8%]
Micro MI	12.3% [11.5% - 13.2%]	2.4% [1.9% - 2.9%]
Micro RI	12.4% [11.5% - 13.2%]	3.5% [3.0% - 4.2%]
ROVER majorité	**10.3%** [9.5% - 11.12%]	**1.8%** [1.4% - 2.3%]
Modèles adaptés sur R et S micro considéré		
Micro LE	11.5% [10.7% - 12.4%]	2.4% [2.0% - 3.0%]
Micro MI	11.2% [10.4% - 12.1%]	2.5% [2.1% - 3.1%]
Micro RI	11.3% [10.5% - 12.2%]	3.0% [2.5% - 3.6%]
ROVER majorité	**9.9%** [9.1% - 10.7%]	**1.9%** [1.5% - 2.4%]

Tab. 1: Taux d'erreur-mot obtenus avec l'analyse Aurora pour les différents microphones

Le tableau 1 présente les taux d'erreur-mot obtenus sur les corpus T (numéro de téléphone) et W (mot de commande) en utilisant un modèle adapté avec les enregistrements des trois microphones distants des énoncés des corpus de phrases R et S. L'adaptation des modèles de reconnaissance à l'environnement sonore de l'intérieur de l'habitacle permet de diminuer de plus de 40% le taux d'erreur-mot par rapport au modèle phonétique standard, que ce soit pour le corpus T ou le corpus W.

Le tableau 1 présente aussi les résultats des tests effectués en adaptant les modèles de reconnaissance avec des enregistrements provenant uniquement du microphone considéré. Les enregistrements du microphone MI, par exemple, sont ainsi testés avec un modèle adapté à partir d'énoncés enregistrés par le microphone MI. Pour le corpus T, le taux d'erreur-mot reste proche pour le microphone LE, mais l'adaptation sur un microphone spécifique permet d'améliorer sensiblement le taux d'erreur-mot pour les micros MI et RI. La figure 1 regroupe l'ensemble des résultats obtenus sur le corpus T par adaptation du modèle de reconnaissance sur un ou sur trois microphones (avec les intervalles d'erreur). Sur le corpus W, les écarts obtenus ne sont pas statistiquement représentatifs.

4. Variabilité des erreurs entre les microphones

Les résultats globaux de reconnaissance obtenus pour les trois microphones sont proches les uns des autres mais cachent en fait d'importantes différences. Leur comportements varient suivant les conditions d'enregistrements (tableau 2) : les écarts peuvent être

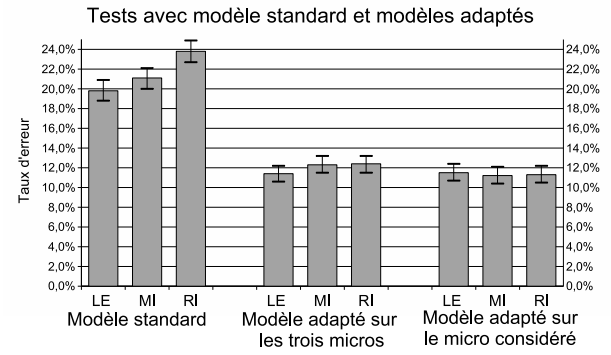

Fig. 1: Evolution du taux d'erreur corpus T

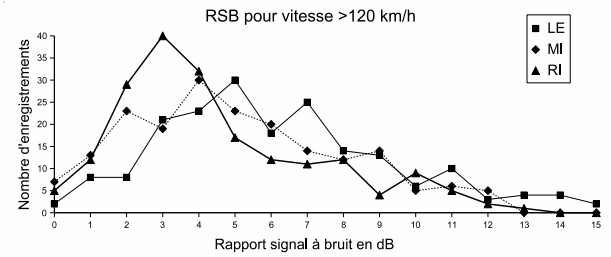

Fig. 2: Répartion des enregistrements par RSB

faibles (0.7% à 30-60 km/h) ou statistiquement importants (6% entre LE et RI à plus de 120 km/h) et certaines conditions impactent plus un microphone que l'autre. Ces différents comportement peuvent être corrélés avec la répartition des rapports signal à bruit pour les trois microphones. Ainsi, le microphone LE est très impacté, du fait de sa position, par l'ouverture de la fenêtre conducteur, et le microphone RI par le bruit à forte vitesse, comme le montre la figure 2 (les rapports signal à bruit sont calculés grâce à un automate de détection bruit-parole [9]).

En outre, ces différents taux d'erreur-mot cachent une importante diversité dans les résultats de reconnaissance obtenus. En effet, les réponses (qu'elles soient correctes ou fausses) sont communes aux trois microphones pour seulement 73.4% des mots des énoncés ">120 km/h" du corpus T et 87.3% des mots des énoncés "Fenêtre ouverte". Dans les autres cas (respectivement 26.6% et 12.7%), les mots des énoncés sont reconnus différemment suivant le microphone d'enregistrement. Le point à ne pas négliger est que parmi ces réponses différentes, certaines sont correctes et n'appartiennent pas obligatoirement au microphone présentant le meilleur taux d'erreur-mot global.

Le tableau 3 montre que sur l'ensemble du corpus T, plus de 7% des mots ne sont reconnus correctement que par un seul microphone si l'on compare deux microphones entre eux (résultats en gras). Pour les microphones LE et RI sur les énoncés ">120 km/h" on atteint 14.3% (9.8% LE correct, RI faux et 4.5% LE faux, RI correct), alors que l'écart entre les taux globaux est de 8% en faveur de LE. Il est donc intéressant de tirer profit de ces écarts entre les microphones et de combiner les résultats de reconnaissance afin obtenir de meilleures performances globales.

	LE	MI	RI
Modèles adaptés sur R et S LE-MI-RI			
30-60 km/h	9.8% [8.9% - 10.8%]	10.5% [9.5% - 11.5%]	9.8% [8.9% - 10.8%]
>120 km/h	16.1% [13.9% - 18.6%]	20.9% [18.5% - 23.6%]	22.1% [19.6% - 24.8%]
Fenêtre fermée	9.3% [8.3% - 10.4%]	10.6% [9.5% - 11.8%]	9.6% [8.6% - 10.8%]
Fenêtre ouverte	12.0% [10.2% - 14.0%]	10.4% [8.7% - 12.3%]	11.0% [9.3% - 13.0%]

Rover			
30-60 km/h	8.6% [7.6% - 9.7%]	>120 km/h	17.0% [14.7% - 19.5%]
Fenêtre fermée	8.6% [7.6% - 9.7%]	Fenêtre ouverte	9.9% [8.3% - 11.9%]

Tab. 2: Différence de comportement entre les microphones suivant les conditions d'enregistrements pour le corpus T (fenêtre ouverte/fermée à 30-60 km/h)

Micro		LE		MI	
		Correct	Faux	Correct	Faux
MI	Correct	84.9%	**3.3%**		
	Faux	**4.2%**	7.6%		
RI	Correct	84.6%	**3.6%**	84.4%	**3.9%**
	Faux	**4.5%**	7.3%	**3.8%**	7.9%

Tab. 3: Différence de reconnaissance suivant le microphone d'enregistrement pour le corpus T (avec modèles de reconnaissance adaptés sur les trois micros)

5. Utilisation conjointe des résultats des trois microphones

Une façon de combiner les résultats provenant des différentes expériences est d'utiliser l'approche ROVER (Recognizer Output Voting Error), développée par NIST [10]. Elle cherche à l'origine à réduire le taux d'erreur-mot de la reconnaissance automatique de la parole en utilisant les différences dans la nature des erreurs faites par différents systèmes de reconnaissance sur un même énoncé. Il s'agit de regrouper les différentes sorties en un seul WTN (word transcription network, voir figure 3) et de comptabiliser les votes reçus par chaque mot présent dans un set de réponse. Cette construction se fait de manière itérative et l'ordre des sorties a une importance. Les expériences ont été menées en utilisant l'ordre LE-MI-RI (choisi pour ses performances sur la globalité du corpus T). L'approche ROVER tient aussi compte des erreurs de suppression/d'insertion, en utilisant un "mot vide" noté @, avec un score de confiance $conf(@)$ défini lors du paramétrage du vote.

Le score final d'un mot w est obtenu à partir de sa fréquence d'apparition $N(w)/Ntotal$ dans le set et son score de confiance $C(w)$ (pour un mot apparaissant plusieurs fois, on peut considérer soit le score maximal, soit une moyenne des scores de confiance). On utilise la formule suivante, avec $0 \leq \alpha \leq 1$:

$$Score(w) = \alpha(\frac{N(w)}{Ntotal}) + (1-\alpha)C(w) \qquad (1)$$

Le défaut majeur de cette approche est que plusieurs mots peuvent se retrouver avec le même nombre de votes, et donc avec un score identique si on ne dispose

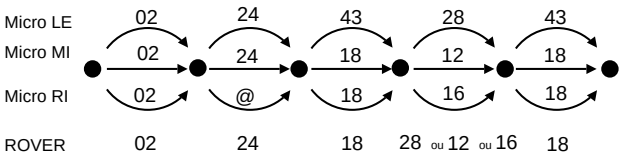

Fig. 3: Word Transcription Network utilisé par Rover

Micro. de test	Rover maj.	Rover RSB
Modèles adaptés sur R et S LE-MI-RI		
Corpus T	10.3% [9.5% - 11.1%]	10.2% [9.4% - 11.0%]
Modèles adaptés sur R et S un seul micro		
Corpus T	9.9% [9.1% - 10.7%]	9.7% [9.0% - 10.5%]

Tab. 4: Taux d'erreur-mot obtenus avec le vote à majorité de ROVER et une pondération liée au RSB

pas de score de confiance. Dans ce cas, le choix entre les prétendants se fait de manière arbitraire.

A la différence d'une utilisation classique de ROVER, nous n'allons pas combiner les sorties de différents systèmes de reconnaissance pour un enregistrement, mais les sorties d'un unique système de reconnaissance pour trois enregistrements d'un même énoncé.

5.1. Vote à majorité

Les tests ont tout d'abord été menés en ne considérant que la fréquence d'apparition du mot ($\alpha = 1$), par un vote à majorité. En utilisant un modèle adapté sur les trois microphones, on obtient des taux d'erreur-mot de 10.3% [9.5% - 11.1%] sur le corpus T et de 1.8% [1.4% - 2.3%] sur le corpus W (voir tableau 1), statistiquement meilleurs que ceux que l'on obtient en ne considérant qu'un seul micro. En utilisant les modèles adaptés sur le microphone considéré, on obtient un taux d'erreur de 9.9% [9.1% - 10.7%] sur le corpus T. Les expériences sur des groupes particuliers d'énoncés donnent des résultats similaires (tableau 2), sauf dans le cas ">120 km/h" où l'approche ROVER donne un résultat moins bon que celui du microphone LE, mais tout de même meilleur que MI et RI. Ce résultat est lié à l'écart important entre le taux du microphone LE et celui de MI et RI.

5.2. Pondération liée au RSB

Pour aller plus loin dans l'utilisation de ROVER et pour pouvoir briser de manière non-arbitraire les égalités lors des votes (2% des énoncés pour le corpus T), nous avons vu qu'il était possible d'utiliser des scores de confiance, compris entre 0 et 1, pour chaque mot (Equation 1). Comme nous ne disposons pas de ces scores, nous avons décidé d'appliquer à la place une pondération liée au rapport signal à bruit (RSB), afin de favoriser, pour chaque énoncé, le microphone proposant le meilleur RSB. L'enregistrement présentant le meilleur RSB se voit assigner une pondération de 0.75, les deux autres recevant 0.5. L'utilisation de ces pondérations et de ROVER, avec des valeurs pour α et $conf(@)$ correctement choisies (par recherche exhaustive par pas de 0.1 sur le corpus T), permet d'obtenir une légère baisse du taux d'erreur, mais qui reste statistiquement insuffisante (voir tableau 4).

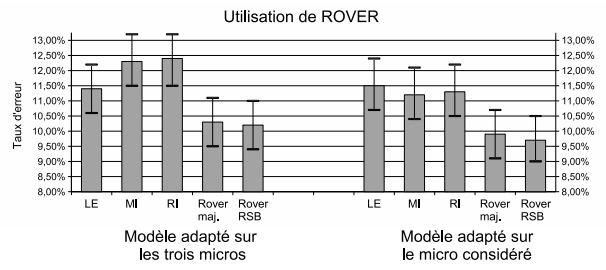

Fig. 4: ROVER améliore le taux d'erreur

6. Conclusion

Dans cet article, nous avons combiné les résultats de reconnaissance des enregistrements d'un même énoncé par plusieurs microphones présents dans un habitacle de voiture. Grâce à l'approche ROVER et aux différences de comportements de ces microphones face à diverses conditions d'enregistrements, nous avons pu obtenir des performances de reconnaissance supérieures à celles obtenues avec un seul microphone. Appliquer cette approche avec un système fournissant des scores de confiance permettrait sans doute d'améliorer davantage le taux d'erreur de la reconnaissance vocale.

Références

[1] Wakita T. Speech-Based Interfaces in Vehicles. *R&D Review of Toyota CRDL*, 39 :1–3, 2004.

[2] M. Westphal and A. Waibel. Towards Spontaneous Speech Recognition for On-Board Car Navigation and Information Systems. *Proceedings of the Eurospeech 99*, 1999.

[3] Villing J. Does a Speech Interface Increase the Safety when Driving a Car. *Dept of Computational Linguistics, Goteborg Univ.*, 2004.

[4] Y. Gong. Speech Recognition in Noisy Environments : A survey. *Speech Communication*, 16 :261–292, April 1995.

[5] Esti document - es 201 108. *STQ : DSR ; Front-end Feature Extraction Algorithm ; Compression Algorithm*, 2000.

[6] http ://parole.loria.fr/. 1998.

[7] L. Delphin-Poulat. *Comparison of Techniques for Environment/Application Adaptation in a Telephony Context. ITRW*, pages 139–142, August 2001.

[8] L. Delphin-Poulat. *Robust Speech Recognition Techniques Evaluation for Telephony Server Based In-car Applications. ICASSP*, pages 65–68, May 2004.

[9] Charlet D. Martin A. and Mauuary L. *Robust Speech/Non-Speech Detection using LDA applied to MFCC. ICASSP*, pages 237–240, May 2001.

[10] J.G. Fiscus. A Post-Processing System to Yield Reduced Word Error Rates : Recognizer Output Voting Error Reduction (ROVER). *Automatic Speech Recognition and Understanding*, pages 347–354, 1997.