Un Système de Détection de la Parole pour les Environnements Perceptifs Multimodaux

Dominique Vaufreydaz, Rémi Emonet, Patrick Reignier

PRIMA - INRIA Rhône-Alpes
Zirst, 655 avenue de d'Europe, Montbonnot, 38334 Saint Ismier cedex, France
{Dominique.Vaufreydaz,Remi.Emonet,Patrick.Reignier}@inrialpes.fr
http://www-prima.inrialpes.fr/

ABSTRACT

In with paper, we address the problem of speech activity detection in multimodal perceptive environments. In such space, one can find many different microphones (lapel, distant or table top). Thus, we need a generic speech activity detector in order to cope with different speech conditions (close-talking, noisy distant speech). Moreover, as the number of microphones in the room can be very high, we also need a very light system.

The speech activity detector presented in this article works efficiently on dozens microphone in parallel. Its accuracy can lead up to ~90% for close-talking microphone and ~85% for distant speech. This system is currently used in several activities in ubiquitous computer research at our laboratory.

1. INTRODUCTION

Dans les recherches menées sur l'ordinateur évanescent (*ubiquitous computing*), le principe de base est la plus grande transparence possible de l'ordinateur dans un système d'interaction homme/machine. Les entrées/sorties standards des systèmes informatiques (clavier, souris, écran, etc.) sont alors remplacés par des modalités moins intrusives pour l'utilisateur comme par exemple la voix ou l'observation par des caméras. De nombreux laboratoires de recherche se sont dotés de lieux comprenant une grande quantité de capteurs (caméras, microphones, etc.) et de logiciels de perception (suivi vidéo 2D/3D de cibles, reconnaissance de la parole, …) pour mener à bien leurs recherches dans ce domaine. Le but est de permettre au système informatique de percevoir ce que les utilisateurs sont en train de dire ou de faire, de pouvoir comprendre ce qu'ils font et d'agir en conséquence. Ces lieux sont ce que l'on nomme des environnements perceptifs multimodaux.

Dans ce genre de systèmes, il est évident que les capteurs sonores, des plus simples aux plus évolués, ont un très grand rôle à jouer sachant que la parole est l'un des modes de communication les plus naturels de l'être humain et que certains sons sont très révélateurs de l'activité humaine, c'est-à-dire de ce qu'est en train de faire l'utilisateur. Ainsi, ces environnements perceptifs sont souvent équipés de système de reconnaissance de la parole, de localisation acoustique ou de détection de la parole comme dans le projet CHIL [1] par exemple. Ces environnements supposent donc de pouvoir traiter en parallèle un très grand nombre de microphones, de plusieurs types en même et tout ceci en temps réel.

C'est dans ce cadre et au sein du projet CHIL[1] que nous développons notre système de détection de la parole. Même si de nombreuses recherches ont déjà été menées dans le passé sur le sujet et de nombreuses approches déjà proposées (comme par exemple [2] [3]), le problème n'est pour l'instant pas entièrement résolu. De plus, nous nous posons des contraintes supplémentaires par rapport à la majorité des systèmes. Nous ne souhaitons pas devoir refaire un entrainement de notre système à chaque fois que les conditions d'utilisation changent et nous souhaitons que celui-ci soit le plus léger possible.

Nous commencerons cet article par la description de notre système. Nous présenterons ensuite les évaluations qui ont été conduites et nous finirons en concluant par les voies d'amélioration qui nous sont ouvertes.

2. DESCRIPTION DU SYSTEME

Le système de détection de la parole que nous avons développé est basé sur une approche à composants tel que nous pouvons le voir sur la Figure 1. Il intègre différents sous-systèmes : un détecteur utilisant les variations de l'énergie du signal, un classifieur basique et un réseau de neurone entraîné pour reconnaître des segments de parole voisés comme les voyelles par exemple. A tout moment, i.e. pour chaque trame de signal, chaque sous-système donne une réponse indiquant si le signal en entrée est de la parole ou non. Ensuite, un ensemble de règles définies manuellement (indépendantes du lieux et du locuteur) intègre ces résultats pour donner la réponse finale du système : parole ou non parole.

[1] http://chil.server.de/

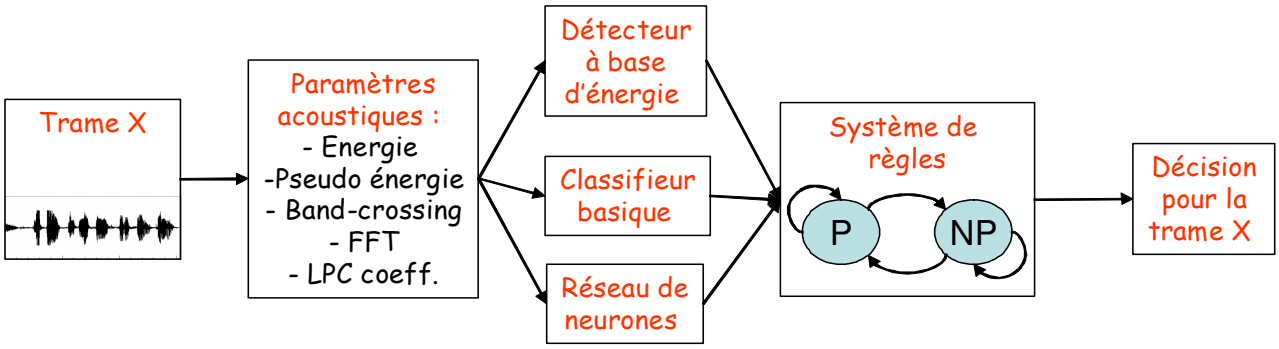

Figure 1 : Description schématique du système

2.1. Détecteur à base d'énergie

Le détecteur à base d'énergie utilise comme entrée la pseudo énergie du signal (on ne somme que la valeur absolue des échantillons) pour déterminer les variations du signal d'entrée. Il fonctionne en utilisant 2 couples de seuils : *TimeOn*/*EnergyOn* et *TimeOff*/*EnergyOff*. Simplement, si l'énergie du signal de parole reste au dessus d'*EnergyOn* pendant plus de *TimeOn*, le détecteur envoie *START_SPEAKING*. A contrario, si par la suite l'énergie passe en dessous d'*EnergyOff* pendant au moins *TimeOff*, le résultat est *STOP_SPEAKING*. Durant les périodes stables, les réponses possibles sont *STILL_SPEAK* et *NOT_SPEAK*.

Dans sa dernière évolution, les seuils d'énergie sont ajustés en ligne pour coller le plus possible à la réalité du signal. La décision finale prise par le système de règles est utilisé comme entrée pour re-estimer dynamiquement *EnergyOn* et *EnergyOff*. Pour éviter tout les problèmes d'*outliers*, le système utilise une médiane plutôt qu'une moyenne. Les 30 dernières secondes des périodes stables de parole et de non parole sont utilisées pour ce calcul.

2.2. Classifieur basique

Ce classifieur est dédié à l'identification de 3 différentes classes de son : les fricatives, les sons basses fréquences comme les ventilateurs d'ordinateur ou d'air conditionné, et les autres sons. Pour cela, la première étape est de calculer le spectre du signal avec un fenêtrage de *Hamming* suivi d'une transformée de Fourier (FFT). Ensuite, la bande de fréquence de 1 à 8000 Hz est découpée régulièrement en 5 sous-bandes de taille identique. Pour chacune d'elles, le classifieur calcule l'énergie. Ce composant ne fonctionne qu'avec des signaux enregistrés à 16000 hertz ou plus. Dans ce dernier cas, les fréquences au dessus de 8Khz sont ignorées.

Avec les 5 valeurs d'énergies de chacune des sous-bandes, ce module peut décider de quelle classe il s'agit en appliquant dans l'ordre les règles suivantes:

- si plus de 90% de l'énergie totale du signal est concentrée dans les 2 bandes basses, c'est un son basses fréquences
- si l'énergie des 2 bandes de fréquence les plus basses est inférieure à toutes les autres bandes, le son est une fricative
- dans tous les autres cas, le son est inconnu.

2.3. Réseau de neurones

Le réseau de neurone est le seul composant qui nécessite un entraînement. Cependant, comme nous ne voulons pas être dépendant d'un microphone ou d'un type de signal, nous avons pris pour l'apprentissage des énoncés du corpus Bref80 [4]. Ces énoncés ont été sélectionnés au hasard mais de façon à garantir la parité entre les voix masculine et féminine. De plus, les étiquettes acoustiques utilisées pour déterminer les portions de signal de parole et de non parole ne sont pas celles fournies avec le corpus. Elles ont été calculées par le système de reconnaissance du français RAPHAEL [5] car les frontières de mots ainsi fixées sont plus proches de celles que nous souhaitons détecter. Le corpus d'entraînement fait environ 1 heure de parole (~31 minutes pour les femmes).

Le réseau de neurone est multicouche avec 2 couches cachées. Il utilise un ensemble de coefficients acoustiques extraits de la trame d'entrée :

- *Zero Crossing* : le nombre de fois où le signal passe d'une valeur positive à une valeur négative et vice versa. En fait, nous utilisons une variante appelée *Band Crossing* [6] qui ne considère pas les faibles oscillations autours de 0.
- *Energie* : la somme des carrés des échantillons
- *16 coefficients prédictifs* : nous utilisons la prédiction linéaire (*Linear Predictive Coding* **[7]**) par auto-corrélation avec la récursion de Levinson-Durbin pour les calculer.

Dans toutes nos expérimentations, nous utilisons l'apprentissage *a priori* que nous avons présenté. En effet, nous souhaitons produire un système générique qui n'est pas dépendant de données d'apprentissage spécifiques pour fonctionner.

3. EVALUATION

3.1. Implémentation

L'implémentation de tous les composants est faite en C++ et fonctionne sur plusieurs plateformes (Windows et Linux). Comme nous l'avons déjà dit, le système peut traiter des signaux de 16 à 44 Khz. A 16000 Hz, la taille des trames en entrée est de 256 échantillons. Le système prend une décision toutes les 16 ms.

L'un de nos objectifs étant de fonctionner en temps réel dans des environnements perceptifs sur de multiples canaux, notre détecteur doit être le plus léger possible. Sur un ordinateur standard du marché, et sur un signal échantillonné à 16Khz en 16 bits, le temps de calcul pour 1 seconde de signal est de 0,0076 seconde. En théorie, notre système serait donc capable de traiter plus de 130 canaux en parallèle. En pratique, si l'on considère les mécanismes de gestion mémoire et d'ordonnancement des systèmes d'exploitation, nous pouvons tabler sur une capacité de traitement d'environ 50 canaux.

3.2. Données d'évaluation

Les données de test que nous allons utiliser sont extraites des évaluations conduites dans le projet CHIL. Les données acoustiques de test sont extraites de séminaires qui ont été enregistrés dans la « CHIL-room » du laboratoire ISL[2]. Cette salle mesure 5,9 m par 7,1 m. Elle est utilisée comme salle de travail par les étudiants pour leurs recherches. Elle est équipée de nombreux capteurs (caméras et microphones). Les données ont été recueillies simultanément par le bais d'un microphone cravate Sennheiser et par un tableau de microphones distants placé à environ 4,5 m en face du séminariste. Nous obtenons donc 2 conditions d'expérimentation que nous nommerons respectivement *Close-Talking* et *Far-Field*.

Le corpus de test se compose de 7 séminaires donc de 7 voix différentes. Les séminaires ont été donnés en anglais par des étudiants dont ce n'est pas la langue maternelle. Ils ont été manuellement transcrits pour permettre l'évaluation des systèmes. Pour chacun d'entres-eux, nous avons 4 segments de 5 minutes chacun. 2 de ces segments servent pour le test, les 2 autres peuvent par exemple servir à l'entraînement des systèmes. Nous n'utilisons pas ces segments, notre système a été utilisé directement sans une phase préalable d'adaptation aux données.

Au total, nous avons donc un corpus 70 minutes de test pour l'évaluation de notre système.

[2] http://isl.ira.uka.de/

3.3. Métriques

Dans le cadre de ces évaluations, au sein du projet CHIL, nous nous sommes mis d'accord sur l'utilisation des métriques :

- *Mismatch Rate (MR)* : durée des décisions incorrectes / durée totale des phrases. C'est le taux d'erreur du système.
- *Speech Detection Error Rate (SDER)* : durée des erreurs dans les segments de parole / durée totale des segments de parole. C'est le taux d'erreur du système dans les portions de parole.
- *Non-speech Detection Error Rate (NDER)* : idem que la précédente mais pour les segments non-parole.

Dans les résultats présentés ci-après, les résultats sont moyennés pour tous les énoncés.

3.4. Résultats

Nous allons présenter 2 catégories de résultats. La première est le système sans adaptation automatique en ligne du détecteur à base d'énergie. Nous trouverons les résultats dans la Table 1. La Table 2 montre le gain sur le même corpus en utilisant cette adaptation.

Dans la première expérience, les seuils d'énergie utilisés sont ceux définis empiriquement lors de précédents projets (NESPOLE ! [8] et FAME [9]).

Table 1 : Résultats obtenus sans adaptation en ligne.

	MR [%]	**SDER** [%]	**NDER** [%]
Close-talking	11.46	3.55	37.72
Far-field	14.83	4.12	58.01

Les performances de notre système sont dans la moyenne de ceux de ceux proposés par les autres partenaires de CHIL. Nous voyons dans le tableau que dans les 2 conditions, la majorité des erreurs de notre détecteur se situent dans les segments non-parole du signal. Si l'on regarde en détail les erreurs commises, elles sont de plusieurs ordres. Premièrement, notre système possède une trop grande latence pour changer d'état. La majorité des erreurs proviennent des transitions parole/non parole. Ce problème survient probablement à cause des valeurs des seuils d'énergie qui ne sont pas adaptés à notre corpus. L'adaptation en ligne de ces paramètres (expérimentation suivante) devrait répondre à cette difficulté.

Les autres sources d'erreur sont les bruits ambiants (ordinateur, air conditionné, porte, etc.) et les bruits du locuteur principal (respiration, …) qui sont parfois interprétés par notre détecteur comme des segments de parole. Si l'on considère que BREF, notre corpus d'entraînement, a été enregistré en environnement calme sans bruit, cela n'est donc pas surprenant.

Dans l'expérimentation suivante, les seuils *EnergieOn* et *EnergieOff* sont recalculés en permanences. Leurs

valeurs initiales sont les mêmes que dans l'évaluation précédente.

Table 2 : Résultats obtenus avec adaptation en ligne.

	MR [%]	**SDER [%]**	**NDER [%]**
Close-talking	10.78	5.83	27.90
Far-field	13.41	12.87	15.40

Le premier résultat de notre système d'adaptation est le gain sur la performance globale du système (MR) pour les 2 conditions même si celui-ci n'est pas significatif. Les changements les plus notables concernent la répartition des erreurs dans les segments parole et non-parole. Notre stratégie d'amélioration a donc profité majoritairement aux transitions parole/non-parole comme le confirme l'étude détaillée des résultats. Cela confirme bien l'intérêt de cette adaptation comme nous nous y attendions. Cependant, cela s'est aussi traduit par une hausse des erreurs dans les segments parole. Il nous faudra revoir la méthode de calcul de nos seuils pour proposer une alternative plus performante pour les segments de parole.

4. CONCLUSION ET TRAVAUX FUTURS

Notre système de détection de la parole présente des performances tout à fait convenables pour les utilisation que nous en faisons que se soit dans le cadre du projet CHIL ou d'autres projets comme par exemple [9] ou [10] et cela dans différents lieux sans travail d'adaptation préalable (bureau intelligent, salon de conférence, amphithéâtre, …). Cependant, comme les expérimentations nous l'ont montré, il nous faut encore procéder à quelques améliorations. Dans un premier temps, l'adaptation en ligne des paramètres du détecteur à base d'énergie doit être revue. En effet, même si celle-ci est bénéfique globalement, il n'en reste pas moins qu'elle pénalise les résultats dans les segments parole. De plus, nous devons aussi envisager d'étendre l'adaptation aux paramètres temporels pour prendre en compte les différences d'élocution. Nous devons aussi envisager d'étendre l'apprentissage de notre réseau de neurones. Encore une fois, l'idée est de garder un apprentissage *a priori*. Nous devons inclure dans cette phase différents types de parole, dans différents environnement et différentes conditions (microphones cravate et distant). Nous obtiendrons ainsi un système générique d'identification de sons voisés. Nous sommes actuellement en train de préparer les corpus qui seront nécessaires pour cette tâche.

À un autre niveau, nous envisageons de remplacer le système de fusion des résultats à base de règles par une approche bayésienne. Ces règles, définies par un expert humain, sont très rigides. Le bayésien pourrait donc le remplacer avantageusement. Cela suppose encore une fois de préparer un corpus pour l'apprentissage.

Enfin la dernière piste que nous envisageons est l'utilisation de toutes les données à notre disposition dans notre environnement perceptif multimodal. En effet, l'utilisation des informations visuelles (comme par exemple, des caractéristiques extraites des visages) pourrait nous permettre d'améliorer les performances de notre détecteur de parole.

BIBLIOGRAPHIE

[1] D. Macho, J. Padrell, A. Abad, C. Nadeu, J. Hernando, J. McDonough, M. Wolfel, U. Klee, M. Omologo, A. Brutti, P. Svaizer, G. Potamianos, S.M. Chu. Automatic Speech Activity Detection, Source Localization, and Speech Recognition on the Chil Seminar Corpus. In *IEEE International Conference on Multimedia & Expo*, janvier 2005..

[2] J. Ramirez, J. Segura, C. Benitez, A. de la Torre, A. Rubio. Efficient voice activity detection algorithms using long-term speech information. *Eurospeech'97*, pages, 1997.

[3] A. Martin, D. Charlet, L. Mauuary. RobustSpeech/Non-Speech Detection Using LDA Applied to MFCC. In *Proc. ICASSP*, vol. 1, 237-240, Salt Lake City, mai 2001.

[4] L. Lamel, J.L. Gauvain, M. Eskenazi, BREF, a large vocabulary spoken corpus for French. In *Proc Eurospeech'91*, Gênes (Italie), 1991.

[5] D. Vaufreydaz, Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue, *thèse en informatique de l'Université Joseph Fourier*, Grenoble (France), 226 pages, janvier 2002

[6] J. Taboada, S. Feijoo, R. Balsa, C. Hernandez, Explicit estimation of speech boundaries, *IEEE Proc. Sci. Meas. Technol.*, vol. 141, pp. 153-159, 1994

[7] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall PTR, ISBN 0-130-15157-2, 1993.

[8] F. Metze, J. Mc Donough, H. Soltau, A. Waibel, A. Lavie, S. Burger, C. Langley, L. Levin, T. Schultz, F. Pianesi, R. Cattoni, G. Lazzari, N. Mana, E. Pianta, L. Besacier, H. Blanchon, D. Vaufreydaz, L. Taddei, The Nespole! Speech-to-Speech Translation System, *Human Language Technologies 2002*, San Diego - California (USA), 6 pages, mars 2002.

[9] F. Metze, P. Gieselmann, H. Holzapfel, T. Kluge, I. Rogina, A. Waibel, M. Wolfel J. Crowley, P. Reignier, D. Vaufreydaz F. Berard, B. Cohen, J. Coutaz, S. Rouillard V. Arranz, M. Bertran,, H. Rodriguez, The "FAME" Interactive Space, *2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Edinburgh - UK, 4 pages, février 2005.

[10] O. Brdiczka, J. Maisonnasse, P. Reignier, Automatic Detection of Interaction Groups. In *Proc. Int'l Conf. Multimodal Interfaces*, octobre 2005.