Caractérisation et détection de parole spontanée dans de larges collections de documents audio

Vincent Jousse[†], *Yannick Estève*[†], *Frédéric Béchet*[§], *Thierry Bazillon*[†], *Georges Linares*[§]

LIA[§], Avignon - LIUM[†], Le Mans
vincent.jousse@lium.univ-lemans.fr

ABSTRACT

Processing spontaneous speech is one of the many challenges that Automatic Speech Recognition (ASR) systems have to deal with. The main evidences characterizing spontaneous speech are disfluencies (filled pause, repetition, repair and false start) and many studies have focused on the detection and the correction of these disfluencies. In this study we define *spontaneous speech* as *unprepared speech*, in opposition to *prepared speech* where utterances contain well-formed sentences close to those that can be found in written documents. Disfluencies are of course very good indicators of *unprepared speech*, however they are not the only ones : ungrammaticality and language register are also important as well as prosodic patterns. This paper proposes a set of acoustic and linguistic features that can be used for characterizing and detecting spontaneous speech segments from large audio databases. To better define this notion of unprepared speech, a set of speech segments representing an 11 hour corpus (French Broadcast News) has been manually labelled according to a level of spontaneity. We present an evaluation of our features on this corpus and describe the correlation between the Word-Error-Rate obtained by a state-of-the-art ASR decoder on this BN corpus and the level of spontaneity.
Keywords: spontaneous speech characterization, spontaneous speech detection, automatic speech recognition

1. Introduction

L'extraction d'information à partir de larges collections d'enregistrements audio nécessite d'établir la structure des documents ainsi que leur contenu linguistique. Par exemple, cela peut consister à ajouter les ponctuations et les limites de phrases aux transcriptions automatiques. Ce processus de segmentation est très important pour nombre de tâches comme le résumé de discours, la traduction ou encore la tâche de *distillation* comme définie dans le projet GALE [1]. Structurer la transcription automatique est un réel défi lorsque l'on traite de la parole spontanée tellement les disfluences et l'agrammaticalité caractérisent ce type de parole.

La parole spontanée est présente dans les journaux d'information sous plusieurs formes : interviews, débats, dialogues, etc. Il ne fait aucun doute que ce qui caractérise le plus la parole spontanée sont les disfluences (les morphèmes spécifiques comme *euh*, les répétitions, les corrections et les faux départs) et plusieurs études se sont concentrées sur la détection et la correction de ces disfluences [2, 3] comme dans la récente évaluation *NIST Rich Transcription Fall 2004*. Toutes ces études montrent une importante baisse de performances entre les résultats obtenus sur des transcriptions de référence et ceux obtenus sur des transcriptions automatiques. En effet, les systèmes de RAP se situant à l'état de l'art en la matière ont des taux d'erreur mot élevés lorsqu'ils transcrivent des données susceptibles de contenir beaucoup de parole spontanée comme de la parole conversationnelle ou des enregistrements de réunions. Un des objectifs de cette étude est d'illustrer le mieux possible ce lien entre le taux d'erreur mot et la parole spontanée.

En plus des disfluences, la parole spontanée est aussi caractérisée par son agrammaticalité et un registre de langage différent de celui qui peut être trouvé dans les textes écrits [4]. En fonction du locuteur, l'état émotionnel et le contexte, le langage utilisé peut être très différent. Dans cette étude nous définissons *la parole spontanée* comme de *la parole non préparée*, en opposition à *la parole préparée* qui se rapproche des phrases bien formées que l'on peut trouver dans des documents écrits. Nous proposons un ensemble de caractéristiques acoustiques et linguistiques pour caractériser *la parole non préparée*. La pertinence de ces caractéristiques est estimée sur un corpus de 11 heures (journaux d'information français) étiqueté manuellement selon un degré de spontanéité sur une échelle de 1 (propre, parole préparée) à 10 (discours haché, à la limite du compréhensible). Nous présentons une évaluation de ces caractéristiques sur ce corpus et décrivons la corrélation entre le taux d'erreur mot obtenu par un système de RAP à l'état de l'art sur ce corpus de journaux d'information et le degré de spontanéité.

2. Caractérisation de la parole spontanée

2.1. Degrés de spontanéité

En définissant la parole spontanée comme de *la parole non préparée*, nous suivons la définition proposée par [5] qui définit un énoncé spontané comme : "un énoncé perçu et conçu au fil de son énonciation". Cette définition illustre la subjectivité de ce classement de la parole en préparée ou spontanée. Idéalement, pour annoter un corpus de parole avec des étiquettes représentant la spontanéité de chaque segment, nous

devrions demander à chaque locuteur d'annoter ses propres énoncés. Ce n'est bien sûr pas faisable, nous avons cependant suivi cette définition en réalisant un protocole d'annotation basé sur la perception du *degré de spontanéité* par un juge humain pour chaque segment donné. Notre approche a été d'étiqueter manuellement les segments d'un corpus de parole avec un ensemble de dix étiquettes correspondant chacune à un degré de spontanéité : le degré 1 correspond à de la parole préparée, assimilable à de la parole lue, et le degré 10 correspond à de la parole très disfluente, presque incompréhensible.

Cette approche nous permet de choisir subjectivement où se place la limite entre parole préparée et spontanée. Dans les expériences nous considérons trois classes : *parole préparée* correspondant aux niveaux de 1 à 3 ; *parole peu spontanée* correspondant aux niveaux de 4 à 6, et *parole très spontanée* correspondant au niveau 7 et plus.

Deux juges humains ont annoté un corpus de parole en écoutant les enregistrements audio. Le corpus a été coupé en segments homogènes (en terme de conditions acoustiques et de locuteurs) grâce à un système de segmentation et de classification en locuteur se situant à l'état de l'art en la matière [6]. Ces segments ne durent pas plus de 20 secondes. Aucune transcription n'a été donnée aux annotateurs. L'accord inter annotateur concernant les degrés de spontanéité a été vérifié sur un corpus d'une heure de journaux d'information. Ensuite, ils ont annoté le reste du corpus séparément. Un des problèmes rencontrés était que des segments de parole spontanée peuvent apparaître n'importe où, pas seulement dans de la parole conversationnelle, mais aussi au milieu d'énoncés très *propres*. De la même manière de la parole conversationnelle peut contenir des segments qui peuvent être considérés comme de la parole préparée. Afin de prendre ces phénomènes en compte, nous avons décidé d'évaluer chaque segment de manière indépendante : un segment de parole spontanée peut être entouré de multiples segments de parole préparée.

Le corpus obtenu après ce processus d'étiquetage est constitué de 11 fichiers contenant des données de journaux d'information français provenant de cinq radios différentes (France Culture, France Inter, France Info, Radio Classique, RFI). Ces fichiers ont été choisis en fonction de leur possibilité de contenir de la parole spontanée en fonction de l'émission. La durée totale est de 11h37 pour un total de 2899 segments (après la suppression des segments ne contenant pas de parole : musique, jingles, ...). Parmi ces segments, 1876 ont été annotés comme étant de *la parole préparée*, 842 comme étant de *la parole peu spontanée* et 203 comme étant de la parole *très spontanée*.

Afin dévaluer l'accord inter annotateur pour cette tâche spécifique, nous avons calculé le coefficient Kappa de cet accord [7] obtenu sur un autre corpus d'une heure de journaux d'information non inclus dans les données décrites auparavant. Le coefficient obtenu sur l'étiquetage en *parole préparée*, *parole peu spontanée* ou *parole très spontanée* est très haut : 0,852. Une valeur supérieur 0,8 est habituellement considérée comme excellente [8].

Parallèlement à cette annotation subjective du corpus, nous proposons par la suite certaines caractéristiques utilisées pour décrire les segments de parole, pertinentes pour caractériser leur spontanéité, extractibles par un système de transcription automatique, et sur lesquelles un système de classification automatique peut être entraîné sur notre corpus annoté. Ce problème a récemment été étudié comme une tâche spécifique de la campagne d'évaluation *Rich Transcription Fall 2004* destinée à détecter les disfluences de la parole. Certaines approches n'utilisent que des caractéristiques linguistiques [3], d'autres des caractéristiques linguistiques et prosodiques [9], ou encore les caractéristiques linguistiques et des caractéristiques plus générales [10].

Dans cet article nous utilisons deux ensembles de caractéristiques : les caractéristiques acoustiques relatives à la prosodie, et les caractéristiques linguistiques relatives au contenu lexical et syntaxique des segments. Nous combinons les deux afin de caractériser le degré de spontanéité d'un segment de parole : cette tâche est différente de la tâche de détection de disfluences puisque les segments de parole spontanée ne contiennent pas nécessairement des disfluences. Par exemple, ils peuvent aussi être caractérisés par une forte variation de débit de la parole. Les caractéristiques utilisées dans cette étude sont brièvement présentées dans la partie suivante.

Caractéristiques prosodiques Les caractéristiques prosodiques utilisées concernent la durée des voyelles et le débit phonémique, comme présenté ci-dessous.

Durée : dans la lignée de travaux précédents décrivant le lien entre la prosodie et la parole spontanée [11], nous utilisons deux caractéristiques : la durée des voyelles et l'allongement des syllabes à la fin d'un mot. Cette dernière caractéristique a été proposée dans [12] et associée au concept de *mélisme*. En plus des moyennes des durées, leur variance et écart-type sont aussi prises en compte pour mesurer la dispersion des durées autour de leur moyenne.

Débit phonémique : de précédentes études [12] ont montré la corrélation entre les variations du débit de parole et l'état émotionnel du locuteur. Partant de cette idée nous avons utilisé comme caractéristique une estimation du débit de parole, mot par mot ou par segment de parole, afin d'observer son impact sur la spontanéité de la parole. Nous estimons le débit phonémique de deux manières : la variance du débit phonémique pour chaque mot, et la moyenne du débit phonémique sur le segment entier, ceci en incluant les pauses et les morphèmes spécifiques (*ben*, *euh*).

Caractéristiques linguistiques La principale caractéristique de la parole spontanée est le concept des *disfluences*. Elles peuvent être catégorisées comme des pauses aussi appelées *fillers*, des répétitions, des corrections ou des faux départs. Beaucoup d'études se sont concentrées sur leur description à un niveau acoustique [11] ou lexical [13]. Nous utilisons deux caractéristiques des segments de parole qui les représentent :

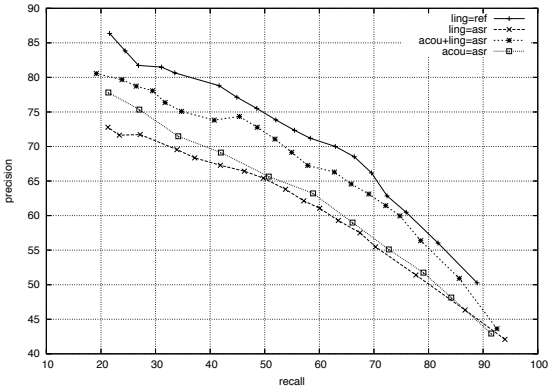

Fig. 1: Performance de la détection de segments spontanés (degré de 4 à 10) en fonction d'un seuil sur le score de classification

– morphèmes spécifiques ou *fillers* : le lexique du système de RAP contient plusieurs morphèmes pour représenter les hésitations comme *euh*, *ben* ou *hum*. Leur nombre d'occurrences total dans un segment est la première caractéristique retenue.
– répétitions et faux départs : nous utilisons une caractéristique très simple qui consiste à compter le nombre de répétitions d'unigrammes ou de bigrammes dans un segment.

2.2. Caractéristiques acoustiques et linguistiques

Comme présenté dans [4] sur des données extraites de journaux d'information, la parole spontanée est aussi caractérisée au niveau linguistique par d'autres phénomènes que les hésitations ou les répétitions. L'agrammaticalité et le registre de langage sont aussi très caractéristiques de la parole non préparée. Afin de capturer le lien entre la spontanéité d'une part et le lexical et le syntaxique d'autre part, nous appliquons aux transcriptions des segments audio un processus de *shallow parsing* [14] incluant un étiquetage en parties du discours (POS) et un processus de découpage syntaxique. Nous utilisons les caractéristiques suivantes pour décrire les segments :

– paquets de n-grammes (de 1 à 3-grammes) sur les mots, étiquetage POS et découpage syntaxique en catégories (groupe nominal, groupe prépositionnel) ;
– taille moyenne des découpages syntaxiques du segment.

3. Détection automatique de segments de parole spontanée

Les caractéristiques présentées dans la section précédente ont été évaluées sur notre corpus étiqueté dans un but de classification : étiqueter les segments de parole selon deux types de parole : *parole préparée* (degré 1-3) ou *parole spontanée (non préparée*, degré 4-10). Le classifieur que nous avons utilisé est *Boostexter* qui se base sur l'algorithme AdaBoost [15]. C'est un outil de classification basé sur la méthode de *boosting* des classifieurs *faibles*. Ces classifieurs faibles sont donnés en entrée : ils peuvent être la présence ou l'absence d'un mot spécifique ou n-gramme (pour les caractéristiques linguistiques) ou des valeurs numériques (pour les caractéristiques acoustiques, les disfluences et la moyenne des tailles des découpes syntaxiques). À la fin du processus d'apprentissage, la liste des classifieurs sélectionnés est obtenue ainsi que le poids de chacun d'entre eux dans le calcul du score de classification pour chaque segment à traiter. Le corpus (comme décrit dans 2.1) est constitué de 11 fichiers audio. Pour nos expériences, nous utilisons la méthode du *Leave One Out* : dix fichiers utilisés pour l'apprentissage, un pour l'évaluation ; ce processus est répété jusqu'à ce que tous les fichiers aient été évalués.

parole préparée				
	ling(ref)	ling(rap)	acou(rap)	tout(rap)
Précision	79.2	75.8	75.9	78.3
Rappel	84.9	80.1	81.2	81.7
F-mesure	81.9	77.9	78.5	80.0
parole spontanée				
	ling(ref)	ling(rap)	acou(rap)	tout(rap)
Précision	70.8	62.7	63.9	66.5
Rappel	62.1	56.7	56.3	61.5
F-mesure	66.2	59.5	59.9	63.9
Tout				
	ling(ref)	ling(rap)	acou(rap)	tout(rap)
Correct	76.4	71.4	72.0	74.2

Tab. 1: Précision, rappel et F-mesure pour la classification des segments de paroles en fonction de 2 catégories : *parole préparée* et *parole spontanée*

Le tableau 1 présente les résultats de la détection (précision, rappel et F-mesures) pour les deux étiquettes de spontanéité. Comme nous pouvons le voir, le rappel est meilleur que la précision sur les segments de *parole préparée*, alors que la précision est meilleure sur les segments de *parole spontanée*. C'est pourquoi il est intéressant d'évaluer comment la précision de la détection peut augmenter en utilisant un seuil de décision donné par le classifieur. La courbe de la précision en fonction du rappel est présentée dans la figure 1. Comme nous pouvons le voir la chute de performance entre les transcriptions de référence et les transcriptions automatiques, à cause des erreurs du système de RAP, est compensée par les caractéristiques acoustiques qui sont plus robustes aux erreurs du système de RAP. Une précisions acceptable de 72% peut être obtenue avec un rappel de 50%.

4. Taux d'erreur mot sur les segments de parole en rapport avec leur degré de spontanéité

Cette étude sur la caractérisation des segments de parole avec des degrés de spontanéité peut être utile afin de traiter des segments de parole spontanée avec un système de reconnaissance automatique de la parole (RAP). Comme le montre la figure 2, degré de spontanéité et taux d'erreur mot sont corrélés. En utilisant le système de RAP du LIUM [6] basé sur le décodeur CMU Sphinx, le taux d'erreur global mot obtenu sur

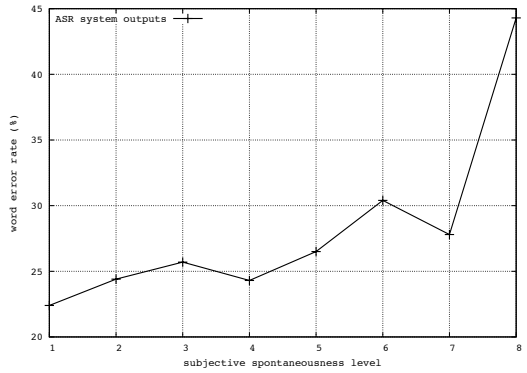

Fig. 2: Taux d'erreur mot en fonction du degré de spontanéité

le corpus entier de 11 heures est de 25,1%. Ce système obtient 23,2% de taux d'erreur mot sur les segments de *parole préparée* (degrés 1-3), alors que le taux d'erreur mot obtenu sur les segments de *parole spontanée* (degrés 4-10) est de 28,3%. Ceci met en évidence que les modèles et les stratégies de recherche utilisés pour traiter la parole préparée ne sont pas suffisants pour traiter la parole spontanée : des caractéristiques spécifiques à la parole spontanée doivent être traitées avec des modèles ou des stratégies de recherche spécifiques (modèle de langage spécifique, dictionnaire de prononciation spécifique, etc.). Détecter ces segments spécifiques pourrait permettre d'utiliser l'approche adéquat.

5. Conclusion

Nous proposons un ensemble de caractéristiques acoustiques et linguistiques qui peuvent être utilisées pour caractériser et détecter les segments de parole spontanée à partir de larges collections de documents audio. Afin de mieux définir cette notion de parole non préparée, un ensemble de segments de parole représentant 11 heures de corpus (journaux d'informations radiophoniques français) a été étiqueté manuellement en fonction d'un degré de spontanéité : la corrélation entre le degré de spontanéité et le taux d'erreur mot obtenu par le système de RAP du LIUM à l'état de l'art en la matière sur ce corpus est présenté. Les caractéristiques acoustiques et linguistiques sont évaluées pour caractériser et détecter les segments de parole spontanée : la combinaison des caractéristiques acoustiques et linguistiques extraites des sorties du système de RAP obtient des performances similaires aux caractéristiques linguistiques extraites des transcriptions de référence seules. Une précision acceptable de 72% avec un rappel de 50% est obtenu pour la détection.

Références

[1] D. Hakkani-Tur and G. Tur. Statistical Sentence Extraction for Information Distillation. *ICASSP 2007*, 4, 2007.

[2] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper. Comparing HMM, Maximum Entropy, and Conditional Random Fields for Disfluency Detection. *InterSpeech 2005*, 2005.

[3] M. Lease, Johnson M., and E. Charniak. Recognizing Disfluencies in Conversational Speech. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5) :1566–1573, 2006.

[4] P.B. de Mareüil, B. Habert, F. Bénard, M. Adda-Decker, C. Barras, G. Adda, and P. Paroubek. A quantitative study of disfluencies in French broadcast interviews. *Proceeding of the workshop Disfluency In Spontaneous Speech (DISS), Aix-en-Provence, France*, 2005.

[5] D. Luzzati. Le fenêtrage syntaxique : une méthode d'analyse et d'évaluation de l'oral spontané. In *MIDL*, Paris, France, 2004.

[6] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. The LIUM speech transcription system : a CMU Sphinx III-based System for French Broadcast News. In *Interspeech 2005*, Lisbon, Portugal, September 2005.

[7] Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 :37–46, 1960.

[8] Barbara Di Eugenio and Michael Glass. The kappa statistic : A second look. *Computational Linguistics*, 30(1) :95–101, 2004.

[9] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5) :1526–1540, 2006.

[10] J.-F. Yeh and C.-H. Wu. Edit Disfluencies Detection and Correction Using a Cleanup Language Model and an Alignment Model. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5) :1574–1583, 2006.

[11] E. Shriberg. Phonetic consequences of speech disfluency. *Proceedings of the International Congress of Phonetic Sciences (ICPhS-99)*, pages 619–622, 1999.

[12] G. Caelen-Haumont. Perlocutory Values and Functions of Melisms in Spontaneous Dialogue. *Proceedings of the 1st International Conference on Speech Prosody, SP*, pages 195–198, 2002.

[13] M.H. Siu and M. Ostendorf. Modeling disfluencies in conversational speech. *ICSLP 1996*, 1, 1996.

[14] Frederic Bechet. LIA_TAGG. http ://www.lia.univ-avignon.fr/chercheurs/bechet/download_fred.html.

[15] Robert E. Schapire and Yoram Singer. BoosTexter : A boosting-based system for text categorization. *Machine Learning*, 39 :135–168, 2000.