Découverte non supervisée de mot(if)s dans le signal de parole

Armando Muscariello, Guillaume Gravier, Frédéric Bimbot

Découverte non supervisée de mot(if)s dans le signal de parole

Armando Muscariello, Guillaume Gravier, Frédéric Bimbot

IRISA UMR 6074 & INRIA Rennes
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
{armando.muscariello,guillaume.gravier,frederic.bimbot}@irisa.fr

ABSTRACT

We propose a method to automatically discover repeating acoustic patterns in speech signals in an unsupervised manner, allowing variability between occurences of a pattern. The resulting patterns, known as audio motifs, are mostly words or sequences of words characterics of the audio content. In this paper, we formalize the problem of motif discovery in speech signals and describe a practical solution using DTW and exploiting the local repetitiveness of motifs. Experimental results on the motif discovery task are provided on a large radio broadcast news corpus. We also propose a refinement of the DTW-based method to account for more variability.
Keywords: motif discovery, audio keyword, unsupervised learning, *data mining*, DTW

1. INTRODUCTION

Dans de nombreuses applications, il est utile de résumer un contenu afin d'en permettre une appréhension rapide. Ainsi, pour les textes, on a généralement recours à quelques mots ou phrases clés tandis qu'en vidéo, on utilise des images clés présentées sous forme d'icônes. En revanche, appréhender un contenu audio directement à partir du signal reste problématique. Dans le cas de contenus oraux, il est évidemment possible d'utiliser une transcription automatique pour se ramener au cas du texte. Mais le processus de transcription automatique est coûteux et parfois peu fiable. La détection de mots clés, ou *word spotting*, présente une alternative intéressante mais limitée à une liste de mots prédéfinis.

Nous étudions ici une approche radicalement différente basée sur la découverte de motifs dans le signal pour faire émerger des icônes sonores correspondant à des mots ou des locutions caractéristiques d'un contenu. La découverte de motifs sonores consiste à détecter à partir du signal des éléments acoustiques réccurents présentant éventuellement un certain degré de variabilité, sans aucune forme de connaissance *a priori*, tant sur le plan acoustique que linguistique. Par exemple, dans le cas de la parole, les mots ou locutions qui se répètent sont des motifs typiques que nous souhaitons voir émerger.

Il convient de bien distinguer la *découverte* de motifs de la *recherche* de motifs. Dans le premier cas, les motifs ne sont pas définis *a priori* tandis que dans le deuxième cas, il s'agira de retrouver un motif connu et défini à l'avance, par exemple par une occurence de référence. Par ailleurs, il est également important de noter que nous souhaitons développer des approches non supervisées dans lesquelles aucune forme d'apprentissage n'intervient. En particulier, nous ne souhaitons utiliser ni modèle de langage, ni modèle acoustique prédéfinis.

Dans le domaine audio, quelques travaux récents s'intéressent au problème de la découverte de motifs. En particulier, Herley propose un algorithme de découverte de motifs sonores quasi invariants pour la découverte d'éléments récurrents (génériques, publicités, *etc.*) dans un flux télévisé [1]. De récents travaux sur la découverte de mots dans le signal de parole relève le défi de la variabilité des motifs [5, 4, 3]. Les approches proposées dans [5] et [4] s'appuient sur un algorithme en deux passes : une première passe vise à détecter des fragments similaires qui sont regroupés dans une passe suivante. Dans [3], nous proposons une approche combinant la stratégie en une passe de [1] avec les méthodes de comparaison de séquences basées sur l'alignement temporel dynamique (DTW). Dans cet article, nous étendons l'approche présentée dans [3] afin d'accroître la robustesse de l'algorithme à la grande variabilité du signal de parole.

Nous formalisons tout d'abord le problème de la découverte de motif avant de détailler l'architecture générale de l'approche proposée. Nous détaillons à la section 4 différentes méthodes pour la comparaison de deux séquences sonores. Les résultats expérimentaux sont rassemblés dans la section 5.

2. FORMALISATION DU PROBLÈME

De manière tout à fait générique, la découverte de motifs consiste à trouver dans un ensemble de données ϕ toutes les paires de segments disjointes, de longueur minimal L_{min}, suffisamment proches. Formellement, on cherche les paires ϕ_a^b, ϕ_c^d telles que

$$H(\phi_a^b, \phi_c^d) < \epsilon \ , \tag{1}$$

où H est une mesure de la distance entre les deux segments, sous les contraintes $b-a > L_{\text{min}}$ et $a < b < c < d$.

Ainsi formulée, la découverte de motifs a pour but de trouver des paires de segments similaires, regroupant ainsi deux occurences d'un même motif. Une étape supplémentaire de *clustering* est ensuite nécessaire pour grouper l'ensemble des occurences d'un motif. Une telle considération nous amène à envisager le problème de découverte de motifs comme un problème de *clustering* se limitant aux portions de signal qui se répètent au moins une fois. Une telle approche s'applique aussi bien lors d'un traitement *a posteriori*, par exemple avec une stratégie multipasse lorsque l'ensemble des données est accessible [5, 4], que pour un

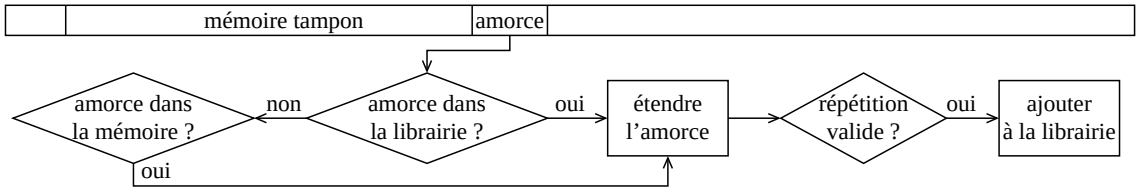

FIG. 1: Schéma de principe de la segmentation du flux et de la recherche pour une amorce donnée.

traitement en flux [1, 3]

Du point de vue conceptuel, nous pouvons décomposer la découverte de motifs en quatre tâches élémentaires : représentation, segmentation, détection et validation. La *représentation* consiste à choisir les descripteurs utilisés pour représenter le signal. La *segmentation* recouvre l'organisation du processus en terme de segmentation des données et d'organisation de la recherche. En effet, une recherche exhaustive de toutes les paires vérifiant (1) n'est bien évidemment pas possible et le recours à une forme de segmentation s'avère indispensable. En particulier, le premier choix à effectuer est celui de la stratégie en une ou plusieurs passes. Enfin, les deux dernières tâches sont directement liées à la comparaison de segments et à la découverte des motifs. La *détection* consiste à identifier les répétitions ϕ_a^b, ϕ_c^d susceptibles de correspondre à deux occurences d'un motif. La *validation* permet par la suite de décider si deux répétitions correspondent en effet à un motif. Cette dernière tâche revient à vérifier (1). Bien que conceptuellement différentes, les tâches de détection et de validation peuvent se résumer en une seule si la même métrique H est utilisée pour les deux.

3. ARCHITECTURE GÉNÉRALE

Nous proposons une approche permettant un traitement en flux des données, dérivée de l'approche ARGOS [1] pour la segmentation. L'idée générale consiste à construire séquentiellement, de manière incrémentale, un catalogue de motifs à partir des données vues comme un flux. Dès lors qu'une nouvelle répétition est trouvée et validée, une nouvelle entrée est créée dans le catalogue, permettant ainsi de retrouver ultérieurement d'autres occurences de ce motif.

La détection des répétitions exploite la notion d'amorce, une amorce correspondant à un segment court, de taille fixé, dans le flux. Une amorce est vue comme un fragment de motif potentiel dont on cherche, dans la phase de détection, à trouver une répétition. Si une répétition de l'amorce est trouvée, on étend alors les segments répétés pour déterminer la répétition la plus longue possible. Cette répétition est ensuite validée comme occurence d'un motif dès lors que les deux segments sont suffisamment proches et insérée dans le catalogue. Afin de limiter le coût calculatoire et de permettre un traitement en flux, la recherche d'une répétition d'une amorce $\phi_t^{t+\delta}$ est limité au passé immédiat $\phi_{t-\Delta}^t$ conservé dans une mémoire tampon. La taille de l'amorce est étroitement liée à la taille minimum des motifs. En effet, l'amorce correspond à un hypothétique fragment de motif et, dans la mesure où l'on cherche une répétition de l'amorce complète, il est important qu'elle ne contienne pas de signal n'appartenant pas au motif lorsque l'amorce est effectivement un fragment de motif. Pour garantir cette propriété, on fixe $\delta = L_{\min}/2$.

Les étapes de l'algorithme sont illustrées par la figure 1. Pour une amorce donnée $\phi_t^{t+\delta}$, on cherche dans un premier temps si cette amorce fait parti d'un motif connu, référencé dans le catalogue, ce dernier étant initialement vide. Si oui, on étend alors l'amorce pour vérifier qu'elle correspond au motif référencé dans le catalogue, remettant à jour le modèle du motif dans le catalogue le cas échéant. Dans nos travaux, le modèle de chaque motif est obtenu par moyennage des occurences trouvées. Si aucun motif du catalogue ne correspond, on cherche dans la mémoire tampon si il existe une répétition de l'amorce de manière à trouver deux occurences candidates pour un nouveau motif par extension de l'amorce. Si un nouveau motif est ainsi découvert, il est ajouté au catalogue après validation. L'algorithme se poursuit ensuite à partir d'une nouvelle amorce localisée soit juste après l'amorce courante si aucun motif n'a été trouvé, soit juste après l'occurence de motif trouvé.

4. DÉTECTION ET VALIDATION

Dans le cadre de segmentation que nous venons de présenter, les tâches de détection et de validation interviennent à deux niveaux, lors de la comparaison avec les entrées du catalogue et lors de la recherche d'une répétition dans la mémoire tampon. Nous décrivons tout d'abord une technique de détection de motifs candidats utilisant une variante segmentale de la technique d'alignement temporel dynamique (DTW) avant de discuter de la validation des répétitions comme occurences d'un motif.

4.1. Détection par DTW segmentale

Rappelons tout d'abord que la phase de détection d'une répétition à partir d'une amorce est un processus en deux étapes. On cherche une répétition de l'amorce – dans le catalogue ou dans la mémoire tampon – avant d'étendre la correspondance de manière à trouver le fragment répété le plus long possible. Nous rappelons ici le principe général de ces deux étapes décrites en détail dans [3].

Considérons une amorce $\phi_t^{t+\delta}$ à rechercher dans un segment χ de longueur $l \gg \delta$. Cette recherche se fait par un algorithme de DTW dans lequel les contraintes de début et fin d'appariement sont relachées, de manière à trouver le fragment de χ apparié au mieux avec l'amorce. Le résultat est un segment χ_s^e tel que sa distance à l'amorce, normalisée par la longueur du chemin d'appariement, notée $D_{\text{DTW}}(\phi_t^{t+\delta}, \chi_s^e)$, est minimum. Les deux segments sont considérés comme une répétition si $D_{\text{DTW}}(\phi_t^{t+\delta}, \chi_s^e) < \epsilon_1$.

La deuxième étape vise à étendre au maximum à gauche et à droite l'appariement existant en s'appuyant sur les points extrêmes. Si l'on prend pour exemple le cas de l'extension à droite (*i.e.*, vers le futur) à partir des deux points $(\chi_e, \phi_{t+\delta})$, on cherche par DTW la meilleure exten-

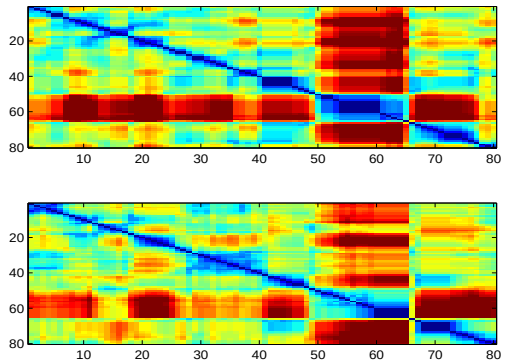

FIG. 2: Exemple de matrices d'autosimilarité d'un motif pour deux locuteurs (masculin en haut, féminin en bas).

sion vers $(\chi_{e+1}, \phi_{t+\delta+1})$, $(\chi_{e+1}, \phi_{t+\delta})$ et $(\chi_e, \phi_{t+\delta+1})$. Le processus d'extension se poursuit tant que D_{DTW} le long du nouvel appariement est inférieure à ϵ_1. Le résultat est une paire de segments, $\phi_{t-\beta_a}^{t+\delta+\alpha_a}$ et $\chi_{s-\beta_b}^{e+\alpha_b}$ telle que $D_{\text{DTW}}(\phi_{t-\beta_a}^{t+\delta+\alpha_a}, \chi_{s-\beta_b}^{e+\alpha_b}) < \epsilon_1$, correspondant à une hypothèse de motif qu'il convient de valider.

L'étape de validation consiste à évaluer (1). La distance D_{DTW} peut être directement utilisée comme métrique H. Cependant, afin d'éviter de valider deux segments différents, cette stratégie requiert un seuil ϵ_1 très petit, limitant ainsi la variabilité tolérée entre occurences d'un motif. En particulier, nous avons observé que cette approche ne permet pas de retrouver des occurences d'un motif par différents locuteurs. Utiliser un seuil ϵ_1 plus élevé autorise une plus grande variabilité au prix d'un nombre plus élevé de fausses détections, c'est-à-dire de détection de répétitions ne correspondant pas à deux occurences d'un motif.

4.2. Validation par matrices d'autosimilarité

Pour pallier au problème précédent, nous proposons une étape de validation exploitant la comparaison de matrices d'autosimilarité. La matrice d'autosimilarité d'une séquence χ_a^b est la matrice carrée $\Phi(\chi_a^b)$ des distances entre points χ_i et χ_j. Clairement, les matrices d'autosimilarité de différentes occurences d'un motif présentent une forte ressemblance visuelle comme illustré par la figure 2. C'est cette ressemblance – interprétable comme une distance entre les autocorrélations plutôt qu'entre les séquences elles-mêmes – que nous souhaitons mesurer et utiliser pour la validation.

La comparaison des matrices d'autosimilarité requiert une normalisation de la longueur des séquences χ_a^b et χ_c^d à comparer, normalisation s'appuyant sur la fonction optimale d'appariement des deux séquences. Étant données les deux séquences normalisées de longueur l, $\tilde{\chi}_a^b$ et $\tilde{\chi}_c^d$, plusieurs métriques sont possibles. La plus simple consiste à prendre la norme l_1 normalisée, soit $D_{\text{SSM}}(\chi_a^b, \chi_c^d) = |\Phi(\tilde{\chi}_a^b) - \Phi(\tilde{\chi}_c^d)|/l^2$. Cette distance reste cependant très dépendante des valeurs absolues des éléments des matrices et ne reflète que peu la similarité visuelle. Afin de prendre en compte la structure spatiale des matrices d'autosimilarité, nous avons recours à une technique basée sur les histogrammes de gradients orientés [2][1]. L'idée générale d'une telle approche est que l'apparence locale d'une matrice d'autosimilarité se caractérise bien par la distribution des gradients d'intensité locaux. Chaque matrice est ainsi transformée en un vecteur de caractéristiques locales, composé des histogrammes des gradients d'intensités pris localement en divers points. La distance entre deux matrices est alors définie comme la norme l_1 entre leurs vecteurs de caractéristiques et notée D'_{SSM}.

Les deux métriques D_{SSM} et D'_{SSM} apportent des informations complémentaires sur la structure des matrices d'autosimilarité. La première mesure directement la différence d'intensité entre les entrées de la matrice. En revanche, la seconde est invariante à l'ajout d'une constante à chaque entrée de la matrice. De plus, en ne se limitant pas à des informations ponctuelles, elle permet de prendre en compte une information plus complexe. En partique, on utilisera donc en parallèle les deux métriques pour valider une répéititon comme occurence d'un motif si $D_{\text{SSM}}(\chi_a^b, \chi_c^d) < \epsilon_2$ et $D'_{\text{SSM}}(\chi_a^b, \chi_c^d) < \epsilon_3$.

5. RÉSULTATS

Nous évaluons tout d'abord l'approche par DTW segmentale pour la découverte de mots dans un flux de parole avant de présenter des résultats préliminaires sur les distances D_{SSM} et D'_{SSM}.

5.1. Découverte de mots dans un flux

Nous avons artificiellement créé un flux de 10 h de signal par concaténation de dix enregistrements d'une heure chacun, dans l'ordre chronologique. Les six premières heures (2 h x 3 chaînes) ont été enregistrées sur une période de 15 jours, les quatre premières correspondant au même jour. Les quatre dernières heures, provenant de 4 chaînes différentes, correspondent à une période de 2 jours, éloignée de 18 mois de la première période. Le choix des données répond à deux considérations majeures. D'une part, on trouve de nombreux mots ou séquences de mots présentant à la fois des répétitions à court terme (au sein d'un reportage par exemple) et à long terme (reportage sur le même sujet mais sur une autre station le même jour ou le lendemain). D'autre part, nous disposons sur ces données d'alignements phonétiques permettant de faire correspondre les motifs découverts au niveau acoustique avec une transcription phonétique.

Dans toutes les expériences, le signal est représenté par des vecteurs de 12 MFCC, plus l'énergie, extraits à une fréquence de 100 trames par seconde.

La qualité des motifs découverts est évaluée au niveau phonétique. Rappelons que le résultat du processus de découverte de motifs est un catalogue de motifs, C_i, chacun caractérisé par ses occurences. La transcription phonétique permet d'associer à chaque occurence j de C_i sa transcription phonétique $C_p(i, j)$. Le motif C_i peut alors être représenté au niveau phonétique par son centroïde, défini comme l'élément $C_p(i, j)$ le plus proche de toutes les occurences du motif. La précision d'un motif correspond alors à la proportion d'occurences suffisamment proche du centroïde. Le rappel est défini par rapport à l'ensemble des chaînes phonétiques suffisamment proches du centroïde de C_i dans la transcription phonétique du flux.

[1]Nous tenons à remercier Émilie Dexter et Patrick Pérez qui ont aimablement mis leurs programmes à notre disposition.

TAB. 1: Précision/Rappel (en %) pour la détection de locutions clés dans un flux de 20 minutes

locution	D_{DTW}	$+D_{\text{SSM}}$	$+D'_{\text{SSM}}$
Jean Marie Le Pen	33 / 59	40 / 59	56 / 59
vingt-et-un avril	18 / 71	22 / 71	43 / 71
extrême droite	17 / 57	25 / 57	67 / 57
France	11 / 43	18 / 39	22 / 35

Pour découvrir des motifs correspondant à des mots ou séquences de mots, nous avons fixé la taille de l'amorce à 0,3 s et celle de la mémoire tampon à 120 s. Le seuil ϵ_1 a été réglé empiriquement de manière à obtenir un bon compromis entre rappel, précision et temps de calcul. Sur les 10h de signal, nous avons trouvé environ 300 motifs, avec une précision de 85 % et un rappel de 25 %. Les motifs trouvés sont donc peu entâchés d'erreurs mais la DTW permet difficilement de grouper des occurences d'un motif qui présente une trop grande variabilité, expliquant ainsi le faible rappel. En particulier, la DTW est très dépendante du locuteur et les occurences d'un même motif par différents locuteurs ne sont pas détectées comme un unique motif mais plutôt comme autant de motifs séparés. Augmenter le seuil ϵ_1 permettrait d'augmenter le rappel au prix d'une forte baisse de la précision. En effet, les motifs dans le catalogue sont représentés par la forme moyenne des occurences trouvées pour ce motif. Augmenter ϵ_1 engendre alors un nombre accru de fausses détections qui viennent déteriorer la représentation des motifs dans le catalogue.

De manière qualitative, les motifs trouvés correspondent principalement à des mots ou des courtes séquences de mots. Par ailleurs, plusieurs motifs sans contenu linguistique sont également trouvés. C'est notamment le cas des inspirations et des *jingles*.

Finalement, il convient de souligner que le temps de calcul pour le traitement des 10h de signal a été d'environ 13h. Même si des optimisations permettrait de décroître de manière significative le temps de calcul, ces chiffres mettent en évidence la difficulté du passage à l'echelle de notre algorithme dans le cas de la découverte de mots. En effet, la taille du catalogue de motifs croît rapidement pour ce type de données, ralentissant ainsi l'algorithme. Ainsi, nous avons mesurer que le temps de traitement en fonction du temps dans le flux est une fonction exponentielle (de la taille du catalogue).

5.2. Utilisation des matrices d'autosimilarité

Avant d'utiliser les métriques D_{SSM} et D'_{SSM} pour la découverte de motifs, nous les avons tout d'abord validé dans un cadre de recherche de motifs connus. Nous avons artificiellement construit un signal de 20 minutes par concaténation de six reportages sur le thème du 21 avril 2002, provenant de radios (et donc de locuteurs) différentes. Quatre locutions clés – Jean-Marie Le Pen, vingt-et-un avril, extrême droite, France –, caractérisées par une occurence de référence chacune, sont recherchées dans les 20 minutes de signal.

Les résultats, en terme de rappel et précision des occurences retrouvées, sont présentés dans le tableau 1. L'algorithme de DTW segmental présenté à la section 4.1 peut être utilisé pour cette recherche (colonne 2), l'occurence de référence du motif à rechercher jouant le rôle d'amorce. Les occurences trouvées pour chaque motif sont ensuite validées en utilisant la distance D_{SSM} (colonne 3), éventuellement complétée par D'_{SSM} (colonne 4). Ces résultats mettent clairement en évidence l'intérêt d'une mesure entre matrices d'autosimilarité pour la validation des motifs, permettant ainsi une amélioration substantielle de la précision pour un rappel constant (à l'exception du motif « *France* », très court). Les occurences trouvées correspondent bien à différents locuteurs, tant masculin que féminin.

Des premières expériences sur l'utilisation des distances entre matrices d'autosimilarité pour la tâche de découverte de motif sur ce court extrait de 20 minutes confirment l'intérêt de ces distances. En utilisant conjointement les deux distances, la précision augmente de 52 % à 66 % et le rappel de 42 % à 51 % par rapport à la seule DTW segmentale. Par ailleurs, l'analyse qualitative des résultats montre que des occurences du motif par différents locuteurs sont retrouvées pour certains motifs, comme « *élevage* »ou « *poisson* ».

6. CONCLUSION

Nous avons proposé une approche pour la découverte non supervisée de motifs sonores dans le signal de parole. La plupart des motifs retrouvés correspondent à des mots ou des séquences courtes de mots qui peuvent être utilisés comme mots clés sonores pour caractériser ou indexer un signal. La méthode utilisant l'alignement temporel dynamique permet de detecter des mots clés avec une bonne précision mais présentent un rappel faible. La combinaison de l'alignement temporel dynamique avec la comparaison des matrices d'autosimilarité permet d'améliorer la découverte de motif au prix d'un effort calculatoire supplémentaire. Ce travail ouvre de nombreuses perspectives, tant pour améliorer la méthode que pour intégrer la découverte de motifs dans des applications d'indexation de documents oraux. En particulier, deux problèmes nous semblent cruciaux. D'une part, le passage à l'échelle reste problématique. Par ailleurs, afin d'utiliser efficacement les motifs découverts, il convient de les caractériser afin de ne conserver que ceux qui décrivent effectivement un contenu linguistique.

RÉFÉRENCES

[1] C. Herley. ARGOS : Automatically extracting repeating objects from multimedia streams. *IEEE Transactions on Multimedia*, 8(1) :115–129, Feb. 2006.

[2] I. Junejo, E. Dexter, I. Laptev, and P. Pérez. Cross-view action recognition from temporal self-similarities. In *Proc. European Conf. on Computer Vision*, pages 293–306, 2008.

[3] A. Muscariello, G. Gravier, and F. Bimbot. Audio keyword extraction by unsupervised word discovery. In *Proc. Interspeech*, pages 2843–2846, 2009.

[4] A. Park and J. R. Glass. Unsupervised pattern discovery in speech. *IEEE Trans. on Acoustic, Speech and Language Processing*, 16(1) :186–197, Jan. 2008.

[5] L. ten Bosch and B. Cranen. A computational model for unsupervised word discovery. In *Proc. Interspeech*, pages 1481–1484, 2007.